Spatial and Temporal Analysis of Volunteered Phenological Observations

FAN SHEN
March, 2012

SUPERVISORS:
Dr. R. Zurita-Milla
MS. Ir. P.W.M. Augustijn

Spatial and Temporal Analysis of Volunteered Phenological Observations

FAN SHEN
Enschede, The Netherlands, March, 2012

Thesis submitted to the Faculty of Geo-Information Science and Earth Observation of the University of Twente in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation.
Specialization: Geo-informatics

SUPERVISORS:
Dr. R. Zurita-Milla
MS. Ir. P.W.M. Augustijn

THESIS ASSESSMENT BOARD:
Dr. Ir. R.A. de By (Chair)
Prof.Dr. C. Robbi Sluter (External Examiner, Federal University of Parana, Brazil)
Dr. R. Zurita-Milla (Member)
MS. Ir. P.W.M. Augustijn (Member)

ABSTRACT

Proper understanding of phenological phenomena, the interrelations among phenological phases of diverse species and how these phenological phenomena are influenced by environmental variables is essential. It is important for understanding critical issues such as climate change. Currently, phenological phenomena studies are extensively based on field experiments but few studies are conducted based on observation information collected by volunteers. Volunteered geographic information (VGI), as a total novel information collection fashion in phenology discipline, starts to play a role in phenological studies. The characteristics of VGI can massive effect to the research and yet potential benefits for phenological revelations cannot be underestimated. One typical characteristic of VGI is spatiotemporal. Due to the complexity of spatial and temporal components, there are a series of challenges in conducting comprehensive analysis of volunteered phenological information.

The aim of this research is to properly analyze phenological phenomena by exploring a spatiotemporal phenological dataset collected by volunteers. The presentation, interaction, interpretation and discovery are combined together for the spatiotemporal phenomena pattern cognition. This research carries out an exploration to integrated methods in spatiotemporal analysis applied on VGI data. The integrated approach are able to: (1) perform identification and the mapping of spatiotemporal patterns present in the phenological dataset with geographic maps, time-series plots, space-time-cube and statistical analysis; (2) discover the space and time synchronous species from 13 common species kinds in the Netherlands with geovisual analytics self-organizing-map and space-time-cube visualization; (3) discover relationships between phenological patterns and common environmental variables (temperatures, evaporation and precipitation) with multivariate analysis in three dimensional presentations and parallel coordinates plots.

Finally, this research shows that “mixed” methods can mitigate each other’s weakness, discover complex spatiotemporal observation pattern and explore the volunteered phenological datasets. For phenologist, this research expands a horizon for phenological phenomena pattern analyst by integrating visual, computational and cartographic methods together to detect and visualize spatiotemporal phenological observation pattern. For researchers in geographic information science, this exploration tells VGI potentials can be revealed.

Key words: VGI, Phenology, Spatiotemporal synchronization, Geovisual analytics, Self-organizing-map (SOM), Multidimensional visualization

ACKNOWLEDGEMENTS

I would like to take this opportunity to appreciate the people who help me during this thesis research.
Firstly, I would like to express my gratitude to my first supervisor Dr. R. Zurita-Milla for the valuable guidance, advice and continuous encouragement contributed to my study.
I also would like to extend my gratitude to my second supervisor Ms. Ir.P.W.M. Augustijn who devoted lots of advice and gave expert suggestions to my thesis.

I am thankful to Chang'an University for giving the opportunity to study in ITC. I really appreciate the support from colleges in this joining programme. Especially, Dr. Xia Li, Ir. M.C. (Kees) Bronsveld, Dr. A.M. Tuladhar, Mr. Xijian Li, Mr. Wei Zhang, Mr. Zheng Zeng and Ms. Xiaojun Yao.
I also deeply appreciate ITC for giving me the great opportunity to study in Netherlands. Life has been really amazing in this busy year.

I am grateful to all Chinese in ITC and appreciate their kind helps. They are Dr. Tiejun Wang, Dr. Xiaogang Ma, Dr. Yali Si, Qiuju Zhang, Xiaojing Wu, Yijian Zeng, Liang Zhou, Xuanmei Fan, Xingping Ye, Biao Xiong, Sudan Xu, Fangfang Chen, Pu Hao, Xuelong Chen, Jing Xiao, Donghai Zheng, Peng Wang, Fangyuan Yu, Ying Jing, Zhi Wang, Zheng Yang, Qifei Han, Chenxiao Tang, Ran Peng, Yang Chen, Chao Yang, Dong Yang, Zhe Kong, Shaoqing Lv and Wei Li. I am feeling home with you fellows.

I want to say thanks to all of my GFM friends. For all the moments we share together, I will never forget in my life.
Great thanks to my colleagues fighting alongside me. They are Mr. Fangning He, Mr. Ding Ma, Mr. Wen Xiao and Miss Bingbing song.

Last but not the least, my sincere appreciations are due to my parents. Your consistent concern and support are parts of the source for my motivation. Without your encouragements, I will not come to this level. Thank you deeply for all you have done for me.

This one is dedicated to you, my beloved parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATION

VGI	*Volunteered geographic information*
PCP	*Parallel coordinates plots*
EDA	*Exploratory data analysis*
SOM	*Self-organizing-map*
STC	*Space-time-cube*
GVA	*Geovisual analytics*

1. INTRODUCTION

1.1. Background and problem statement

In this decade, quantitative spatial data analysis will continue progressing. *"The initial venture into quantitative and mathematical geography led to a wide ranging collection of writings including spatial analysis and pattern analysis"* as indicated by Anselin and Rey (2010). The interest in understanding real-world patterns using methods like statistical analysis, geovisual analytics and other spatial analysis techniques were starting to become gradually apparent. In particular, this research attempts to bring some light on the current understanding of phenological patterns.

The term phenology refers to recurring plant and animal life cycle stages, such as leafing and flowering, maturation of agricultural plants, emergence of insects, or migration of birds. It also refers to the study of these recurring plant and animal life cycle stages, especially their timing and their relationships with weather and climate (NPN, 2011a). The timing of recurring life cycle events differs annually due to environmental factors like temperature or precipitation.

Phenological records can allow scientists to better understand the information from nature events and provide interesting comparisons among years and geographic regions. In recent years, data collected by volunteers has mitigated the problem of having limited collections of phenological observations. In particular, phenology has benefited from what is currently known as volunteered geographical information (VGI). Thus, the novel VGI data source could be a chance for us to expand our horizon. In recent years, there has been a burst of interest to collect geographical information by individuals. The kind of information is named as volunteered geographic information by Goodchild (2007) and the term here refers to geospatial data that is collected by volunteers who are not quite skilled in geography and related areas.

Given a set of spatiotemporal records, one often wants to determine groups with similarities. The spatial or temporal event distribution synchrony such as clustering of the phenological phase becomes the interesting point for researchers that study spatial and temporal analysis. Obtaining and analysing this kind of phenological information is fairly essential (and even plays a role as an indicator) for a number of socio-economic activities, such as agriculture (planting times), natural resources (species suitability), climate change (global warming), public health (hay fever), etc.

In summary, a proper understanding of phenological patterns, the interrelations among phenological phases of the diverse species and how these phenological phenomena are influenced by environmental variables with seasonal and annual variations in weather is significant for understanding critical issues like climate change. But the problem shows up when large amount of multivariate datasets applied in spatiotemporal analysis. It is still difficult to understand the spatiotemporal pattern with conventional techniques. Therefore, this research carries out an exploration to methods in spatiotemporal analysis applied on VGI data. With a Dutch phenological VGI dataset, the main research goal is to properly analysis phenological patterns with spatiotemporal analysis methods.

1.2. Research identification

1.2.1. Main objective

For this research, the overall objective is to properly analyze phenological phenomena by exploring a phenological dataset collected by volunteers.

1.2.2. Sub-objectives

The specific objectives of my research are:

- Identify and map spatiotemporal patterns present in the phenological dataset.
- Study the synchronization of phenological patterns of various species.
- Discover relationships between phenological patterns and environmental factors.

Undoubtedly, the characteristics of VGI, such as large volume, spatiotemporal, affected by human interests, will play a vital role in the data and analysis methods. Therefore these VGI characteristics will be considered in parallel to the research objectives.

1.2.3. Research questions

The questions that are attempted to be answered are as following:

- How can we identify spatiotemporal point patterns of phenological phenomena?
- How can we visualize the phenological patterns of various species?
- How can we compare (e.g. find synchronous groups) phenological patterns?
- How can we represent the relationship from environmental factors to the phenological phenomena pattern?

1.2.4. Innovation aimed at

The novelty of this research comes from the source of data and methods that are applied. Current technology provides us the possibility of accessing VGI dataset. However, it is still lacking powerful and efficient method to explore this novel type of data source. This research will fill in the gap and will allow analyst to explore VGI data through spatiotemporal analysis.

1.2.5. Related work

Researchers are faced with a challenge as there is a lack of systematic theory or established general framework for the growing demand on understanding spatiotemporal patterns. At the meantime, current studies show an increasing trend of interest on spatiotemporal point patterns analysis in a great deal of social-economic regions. For instance, the understanding of spatiotemporal pattern is expanding to criminology. The location and time of crime incidents is studied for socio-economic purposes. Andresen (2009; 2011) successfully tested the criminal pattern in an area-based approach. With the Space-time Cube, Nakaya and Yano (2010) visualized crime cluster in an successful exploratory data-analysis approach.

Furthermore, for multi-dimensional data, the analysis/identification spatiotemporal pattern is often fairly complex. Geovisual analytics, defined by Tomaszewski, et al. (2007), are commonly considered for the analysis/identification of multi-dimensional data and are, therefore, promising for this task. A geovisual analytics tool is tested with a self-organizing network by Ho Van, et al. (2009). While other researchers like Ankerst, et al. applied an enhanced visualization for similarity clustering of multidimensional data (1998).

In a phenological perspective, many researchers believe that, methods for spatiotemporal analysis are becoming increasingly important for ecological studies. It seems that they tend to adopt statistical methods and are focusing various populations of species in the filed experiments when focusing on analysing phenological phenomena synchrony. Bjørnstad and Lambin (1999) put a focus on synchrony analysis of spatial pattern. In Freitas and Bolmgren's research (2008), they have already indicated that the degree of flowering, and also fruiting, synchronization is believed to have ecological and evolutionary relevance at several scales. Researchers like Milla et al. (2010) studied the synchrony pattern of woody plants phenology in the Mediterranean by quantitatively monitoring the phenophase. Similarly, Koenig (2006) examined spatial synchrony in populations of monarch butterflies over the year. Rossi (2011) assessed the synchrony in seed predator Acanthoscelides schrankiae with the reproductive stages of its host plant. While, Keatley, Hudson and Fletcher (2004) studied a long-term (1940–1970) flowering synchrony of Eucalyptus leucoxylon.

Meanwhile, originally developed by Kohonen (1984, 1995), neural network self-organizing maps (SOM) is rarely applied in phenology until Hudson's SOM approach (2011) brings innovative on phenological phenomena synchrony study. Testing self-organizing maps to examine the time series clustering pattern

and represent them in multidimensional representations, the approach inspires many other researchers. The research conducted by Guo et al., combines such computational, visual and cartographical graphs together to detect and visualize multivariate spatial pattern (2005). It shows such mixed methods can mitigate each other's weakness and in an effective way discover complex patterns in large datasets. The view from Andrienko et al. also indicates the fully automatic method is not enough but needs the involvements of analyst's interpretation (2010). In other words, the expanding of whole research horizon on spatiotemporal patterns analysis appears for this research.

1.3. Project setup

1.3.1. Methods

To answer the questions posed in section 1.2.3, the research methods will be applied.

- Literature review: The first task of the research will start from the literature review on novel data source-volunteered geographic information, the characteristics of it. Literature review will also explore the appropriate techniques to identify the spatiotemporal pattern and indicate the task of detecting synchrony of phenology.
- Analyse and prepare the available dataset: Since VGI data set and other environmental data sets will be used in the research. In general the understanding of them is fairly important. In the data preparation, data will then be cleaned to remove error and non-value observations. A subset of data will be done per species and per year for further analysis. After that, new variables may be derived for further processing the data. For instance, a Julian-day attribute can be calculated by using the date of observation.
- Apply the chosen methods to data sets: To identify the spatiotemporal pattern, study the synchronization in phenological patterns of various species and discover relationships between phenological patterns and environmental variables, chosen methods from geovisual analytics and geovisualizations are applied to the available data sets.

1.3.2. Thesis structure

This research thesis contains 6 chapters in total. The background, motivation, problem statement, research objective, research questions and methodology are described in this first introduction chapter.

In chapter 2, literatures are to be reviewed from volunteered geographic information to phenological pattern, techniques in geovisual analytics, geovisualizations and what has to be considered when applying the techniques to volunteered geographic information.

Chapter 3, explanations on applied data and methods will be described. Based on literature review, the suitable methods will be chosen for implementing the research objectives and answer the questions.

In chapter 4, results from applied methods in chapter 3 will be presented. In parallel, discussions for whether the method can fulfil the sub-objectives will be taken and at what level the volunteered geographic information is affected to the results.

Eventually, the research conclusion and recommendation will be drawn in chapter 6.

2. LITERATURE REVIEW

2.1. Introduction

This chapter contains a literature review of the concepts and methods that will be conducted in this research. The goal is to expand the knowledge to meet the implementation need and fulfil the objectives.

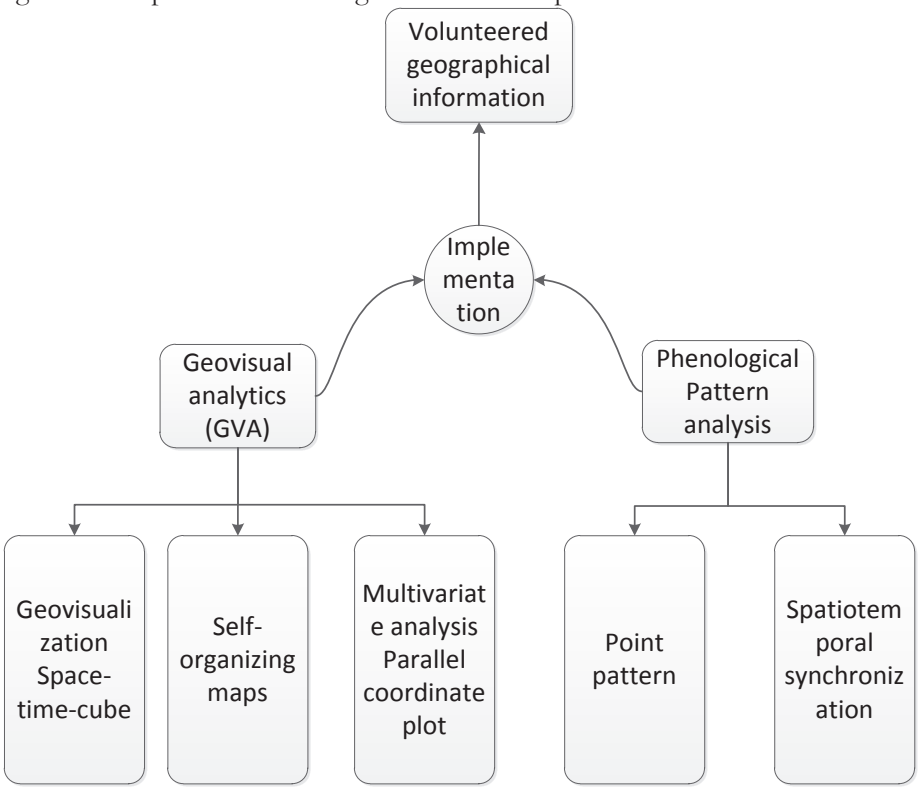


Figure 2-1. Chapter 2 structure chart. The implementation refers to phenological study while geovisual analytics application

Figure 2-1 illustrates the overall structure of this chapter. An overview of volunteered geographical information (VGI) is first conducted. Since in this research, the phenological pattern study is involved and needed to be reviewed. The review is conducted on phenology aspect together with reviewing some species synchronization researches. Then, a brief introduction is given to geovisual analytics. In order to show the detailed specific technology and concepts related to this work. Several subsections are presented. The first two subsections cover the basic theory of geovisualization and provide an introduction to the space-time-cube. The third and fourth subsections are present self-organizing-maps and parallel coordinates plot respectively. The last section defines spatial point pattern analysis and how this analysis can be applied for studying spatiotemporal synchronization.

2.2. Volunteered geographic information

This research will be conducted on a novel data source (volunteered geographical information). In this section, an overview about how volunteered geographic information is played as a role in geo-information science is firstly represented. Typical characteristics of volunteered geographic information and how researchers suggest exploring VGI in studies are also to be reviewed.

In an increasingly fashion, the utilization of concurrent information technologies such as online mapping and mobile devices have enabled possibilities for common people to create and share large volumes of geographical information. As a matter of fact, such personal generated information is usually referred as

volunteered geographic information (VGI). This term was initially coined by Goodchild (2007) and the volunteers here are regarded as ones that are not quite skilled in geography and related fields. In such information acquisition way, the information makes itself as a new role to geo-information science disciplines. The involvement of this information brings not only much added research value but also challenging potentials in nature and social economics regions (Deparday, 2010). For example, the previously scarce information now can be collected by volunteers to backup biodiversity field experiments conducted by ecologist (Vries & Verheul, 2006).

Within geo-information science discipline, VGI is in fact produced with three main purposes, as noted by Elwood (2008). Firstly, VGI is contributed for implementing or updating. This situation happens especially when official data is not sufficient or not updated. Secondly, VGI is produced for replacing or extending the current existing data to avoid restricted finance to users. As noted by Goodchild (2007), the interaction of official data and VGI leads to the creation of a "patchwork" of knowledge. In this "patchwork" the VGI can provide extended knowledge we seek, conforming to typical GIS models. Thirdly, VGI is produced to discover new type of knowledge for both new social and political practices (Deparday, 2010).

Although VGI can be produced in various data types such as conventional spatial points, it is different with conventional data source, such as traditional spatial and attribute data. The reason is that VGI does not refer to a single type of data but rather to data with various characteristics. This data source can have the characteristic of opinion-based, different data capture methods, various accuracy, large volume, point type and spatiotemporal attributed (Deparday, 2010).

Such characteristics are generated during the acquisition of the data. For instance, the data can be spatiotemporal attributed, when volunteers add time and space attributes like exact record time and x, y coordinates. With large number of volunteers' involvement, the collected information can be large volume. As the data capture methods are diverse, the various tools are applied and this can influence the data accuracy. Global positioning system (GPS) devices have emerged into personal mobile devices for common consumers. As one of major tools, it has enabled people to generate geographic data points (Elwood, 2008). The spatial accuracy is far more improved than data collected from Twitter. Information on Twitter is opinion based and data collected from that can be very much biased. Despite this, web 2.0 applications are another way of VGI contribution. It has become widely used in recent crisis events such as the Iranian elections, and earthquake in Haiti and Chile (Tesquet, 2009).

One of most influential VGI characteristics is that it is opinion-based. This leads the relative subjectivity or objectivity of the data. Mentioned by Tulloch (2008) the data can be objective facts. But in the meantime, it is submitted on personal opinions that are subjective. In this respect, VGI can be diverse perspective of facts for individuals.

The following is an example of VGI. For instance, facilitated VGI (F-VGI), coined by Seeger (2008) is one way of data contribution. It is under a planning process and facilitated with digital mapping interfaces. In this case, it can provide facilitated designing professionals with detailed spatial information. This kind of information can be used to create a more informed design solution. Furthermore, another two ways of the data contribution are contributors as a circle of friends, that is common in the social network applications and the unaware contributors (Elwood, 2008).

As potential extended knowledge resources, researchers are arguing with an issue: "in what way can VGI be better explored". Because the influence from VGI characteristics can be led by observations as error or bias. For instance, in ecological region, compared to those professional ecologists, the skill of participants varies in ability, experience and type of training. Thus, a lack of these may definitely lead to error or bias. Studied by Fitzpatrick et al. (2009), the trained volunteers are not as good as professionals at detecting low densities of hemlock wooly adelgids (a bug species). They find that experienced individuals captured small bugs at sites where volunteers failed to do so. And the age of volunteers could also be one of the cause effect elements for research results studied by Delaney et al. (2008).

Preliminary work conducted by Deparday in the new research field (2010) gives us an expanded horizon on VGI exploration issue. In order to deal with challenges and improve the utility of VGI, he investigates the potential of several interactive geovisual analytics and geovisualization techniques. They include filtering, dynamic spatial aggregation, linking and brushing, and tag-based visualizations, together with multivariate analysis to geographic attribute space. Within the space, geographic objects can be located by virtue of their descriptive attributes. Moreover, a review of the challenges and current solutions related to the utilization of VGI is given. Based on this review, a web-based prototype is developed to serve as a platform for the evaluation of selected geovisualization techniques. The prototype is then applied in a series of workshops with large volume citizen-generated data. The result of the case study shows that the implemented geovisualization techniques make users be able to detect relevant subsets of information and get new insights on data. Based on the potential shown by these results, future research direction for applying geovisual analytics techniques to VGI dataset is suggested.

2.3. Pattern of phenology

This research is conducted within a phenological perspective. The term phenology refers to recurring plant and animal life cycle stages. Leafing and flowering, maturation of agricultural plants, emergence of insects and migration of birds are all regarded as parts of the phenological phenomena. It also refers to the study of these recurring plant and animal life cycle stages, especially the timing and their relationships with the environment, weather or climate (NPN, 2011a). The timing of recurring life cycles may diverse annually, due to environmental factors like temperature or precipitation. The variability of this has already received increasing attention as part of the climate change researches (Fitter & Fitter, 2002; Miller-Rushing et al., 2008; Post et al., 2008).

Phenological phenomena are distributed through space and time. As this spatiotemporal pattern (the recurring pattern in location and the timing) of plant, bird, butterfly phenological developments continue to interest ecologists, the synchrony of the phenological pattern has been extensively studied in field experiments (Freitas & Bolmgren, 2008; Keatley, et al., 2004; Koenig, 2006; Milla, et al., 2010). Thus, in phenological studies, the specific spatiotemporal synchronization among species has become the issues that researches tend to reveal. For now, the common sense is the phenological phenomena commence appears in seasons annually. But less is done at larger scale and with novel volunteered geographical information, which can play a facilitated role but yet can be much more potential than we know in pattern recognition. To this extent, the spatiotemporal synchronization pattern analysis is considered challenging to the phenological exploration in this research.

2.4. Geovisual analytics

Since future direction for VGI in geovisual analytics is suggested by Deparday (see section 2.2), a review from general to specific geovisual analytics techniques is conducted in this section. The aim is to find out the appropriate tools for VGI. Firstly, geovisual analytics is presented in detail.

Geovisual analytics (GVA) is an emerging interdisciplinary field that integrates perspectives from visual analytics and geographic information science (growing particularly on work in geovisualization, geospatial semantics and knowledge management, geocomputation, and spatial analysis), noted by Tomaszewski, et al. (2007).

Geovisual analytics tools help to identify relevant spatial information, data, and knowledge. This information can support analytical process involving human vision and cognitions. The supporting knowledge is particularly designed to support for analytical reasoning. Applied in visual interfaces based on the computers, it can provide flexible connections to relevant data. The activities, that geovisual analytics is often conducted into, involve recognizing relevant information in enormous datasets. While using traditional methods, these enormous dataset can make what is relevant difficult to determine. Keim et al. said geovisual analytics is an increasingly important tool for activities ranging from counter-terrorism

and crisis management, through environmental science, to strategic business decision making (2008). Just as the information visualization has changed our view on database, the aim of visual analytics is to make our way of processing information much transparent for an analytic discourse,

The VGI data used in this research is spatiotemporal. This poses serious challenges for the analysis. With respect to the complexity of the geographic space, data source has geographic components requiring the involvement of human analyst's sense associated with place, as noted by Andrienko et al. (2008). Time attribute is complex as well. Though time flows linearly, the phenological phenomena appearing over time is periodically recurring. Phenological phenomena studies have multiple cycles forming in structures, overlapping and interacting in time. The analysts have to have a good understanding of time, but yet can be very hard to convey to the machine (Peuquet, 2002). Thus, data having a temporal component demands the involvement of human interpretation in GVA analysis.

GVA is commonly considered for the spatiotemporal analysis and exploration and therefore are promising for this research. But there are many GVA tools within geo-information science discipline. To study the VGI phenological pattern, the analyst has to understand well the research aim and selects the right method. In this research, to fit for the VGI characteristics of data and fulfil our specific objectives like to study the spatiotemporal synchronization, the following concept or techniques in GVA will be explored.

2.4.1. Geovisualization

As part of GVA, geovisualization is one of widely used techniques. In this subsection, a review on geovisualization is presented to construct a comprehensive theory of geoinformation management and presentation.

Among the geographic information science (GIScience), geovisualization is a field that includes approaches from many disciplines, cartography, scientific visualization, image analysis, information visualization, exploratory data analysis (EDA), to provide theory, methods and tools for the visual exploration, synthesis and presentation of data that contains geographic information (MacEachren & Kraak, 2001). The interactions across various disciplines are fluid, as are the boundaries of these disciplines, claimed by Dykes et al. (2005).

Geovisualizaiton can have roles such as data exploration and experimentation with geographic data. It becomes especially powerful to use geovisualiztion techniques with VGI. In a trend, it can facilitate the multi-dimension explorations in VGI and revealing their relationships. Such work could include the visualization of the social interaction through graph visualization or relationships.

In fact, over the past years, space time activity has attracted considerable research interest in geography. Compared to the recent development of in techniques and applications of multi-dimension geovisualization, Batty and Longley (2003) argued that the development of spatial analysis in multi-dimensional environments was still underdeveloped. One of multi-dimension geovisualization techniques is space-time-cube (STC). This is space time technique showing three axes in a cube. A two dimensional map illustrating the location of phenomenon may not be revealed properly in a true pattern through time. While this three dimensional approach can visualize phenomena that occur repeatedly at the same location and in this case the clusters are able to be noticed. This cube is particularly capable for analysis spatiotemporal attributed data and can be applied in exploratory data analysis. The efforts carried out by Nakaya and Yano (2010) through space-time-cube (STC) geovisualization confirms the technique is promising, because the geovisualization is available in a GIS environment that integrated with various space-time information, such as individual space-time paths and constraints on behaviours (Kwan, 2004). Their example therefore shows large potentials for VGI spatiotemporal data exploration.

This STC in exploratory data analysis is particular matched to spatiotemporal data. But prior to developing visualization based exploratory tools, both clear understandings from data and tasks are essential. Particularly for the VGI data (space and time attributed) that is applied in this research, the spatiotemporal analysis of it can be fairly complex. Andrienko et al. (2005) claim that data visualization methodology has to be carefully designed for exploratory analysis. Because there are two principal aspects may impact and

thus both should be taken into account. They are characteristics of the data to be visualized and the exploratory tasks to be supported. In their research, they demonstrate that different exploratory tasks may be anticipated in three different types of spatiotemporal data cases and different techniques are required to properly support the exploration of the data.

In order to help better understand complex spatiotemporal dataset, descriptive statistics and statistical analysis can be applied to help investigation. From this aspect, various techniques are able to be applied. Lundblad et al. (2008) found multidimensional, multisource, time-varying and geospatial digital information from voyage analysis was represented and able to facilitate the decision-making. The attempt confirmed the descriptive statistics and statistical analysis can help the analysis. Moreover, such fresh insights provide us thinking a helpful statistics perspective for formal analysis.

2.4.2. Self-organizing maps

As our VGI can be large volume, there is one capable GVA technique for this kind of dataset. Self-organizing maps are widely used for complicated datasets with large volumes of data and high dimensionality in attributes. In this subsection, a specific GVA technique self-organizing map is reviewed.

Self-organizing maps (SOMs), also known as kohonen feature map, are a type of artificial neural network introduced by Kohonen (1984, 1995, 2001). This kohonen feature map was first introduced to public by this Finnish professor (University of Helsinki) in 1982. As the simulated human brain learning process, it is one of the most useful neural network types. SOMs are based on competitive learning and they are good tools in exploratory phase of data mining (Vesanto & Alhoniemi, 2000). It has a typical data driven form and data driven content and its system of relationships is generated by the data.

Large volume of SOM applications have been found among various disciplines in an early survey (Kaski, 1997). SOM has especially been used for diverse aims, such as function approximation, data clustering, and dimensional reduction of multidimensional data. There are many variations in type of SOMs and in the context of this research the basic SOM proposed by Kohonen is for application (2001).

SOMs are different from other artificial neural networks. They apply a neighbourhood function to keep the topological properties of the input space. The map self-organizing is that, when the content changes, its form also changes.

As one of the GVA technique, SOMs are applied to many fields. A demonstration of how an SOM works in sea surface temperature patterns is adapted from Liu et al. (2006). The time series data is rearranged in the 2D array. The outcome weight vectors of the SOM nodes are reshaped back into characteristic data patterns. Guo et al. introduces an integrated geographical discovery environment that is able to detect and visualize multivariate spatial patterns for cancer dataset (2005). In their approach, the focuses are on SOMs and a parallel coordinate plot. Giving a comprehensive analysis, G. Andrienko et al. applied a SOM framework for spatiotemporal patterns in 41-years' time series of 7 crime rate attributes in the states of the USA (2010).

While, Hudson et al. (2011) succeeded with testing the capability of SOM correlations to examine clustering patterns. The aim for the approach is that the resultant clusters of species, indicates synchronization of phenological phenomena for the species belonging to the given cluster. Their work indicates the possibility about applying SOM for the synchrony assessment study. This approach gives us large volumes of hints in promising direction of SOM application in the phenology perspective. The SOM could be a very effective and potential technique for exploration to large volumes of VGI data.

2.4.3. Multivariate data visualization

As high volumes of VGI data which may consist or be extended with many attributes, the interests lie in retrieving meaningful information hiding in data. It demands a more powerful, interactive and appropriate visualization methods.

The new developments in GVA technology provide the support to deal with such rich datasets. Information visualization and data mining process help the user to extract, explore and understand the

data at hand. When multivariate representations are generated, they tend to focus on particular attribute domains, with examples including crime statistics (G. Andrienko, et al., 2010), population census data (Skupin & Hagelman, 2005), or medical data from Edsall (2003). Therefore, if VGI is extended with other data source into multivariate datasets, multivariate data analysi should be applied for the exploration to this novel large volume spatiotemporal data source.

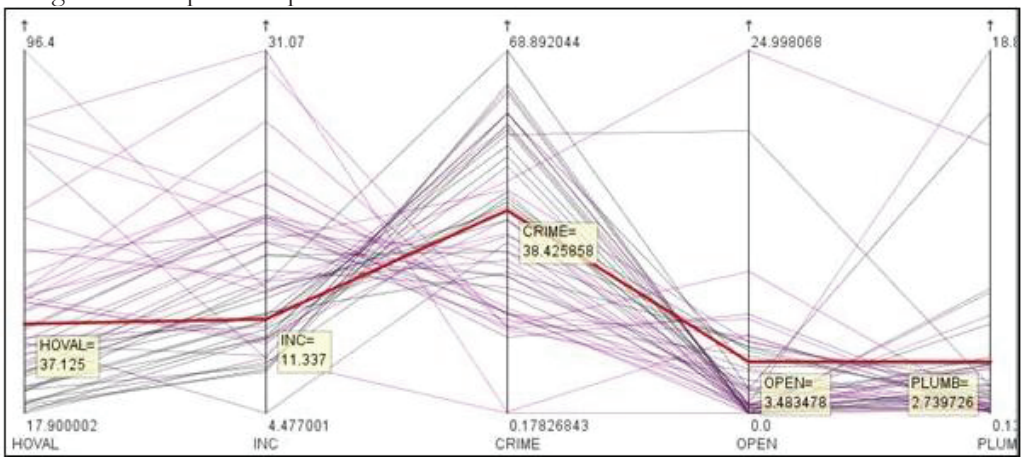


Figure 2-2. Parallel coordinate plots by Smith, et al.,(2008)

Having typical characteristics of multi attributes, VGI need multivariate data visualization. Thus, parallel coordinate plot can be suitable to be conducted. Parallel coordinate plot (PCP) is a graph representation that serves well for depiction of data with high-dimensionality in attributes. The PCP environment represents the geographic objects as a series of continuous lines intersecting a number of vertical lines. The maximum and minimum values of variables are positioned at the top and bottom of vertical lines and the horizontal lines cross the vertical units relative to maximum and minimum values (Kraak & Ormeling, 2010). In Figure 2-2, just like Edsall noted, "the result is a multivariate signature for each observation and a visual representation of relationships among many variables" (2003).

In this typical parallel coordinate plot example illustrated in figure 2-2, five variables are included: house values, income levels, crimes recorded (residential burglaries and vehicle thefts), open space and the percentage of housing with deficient plumbing. The individual variable can be displayed from minimum to the maximum through vertical scale. Linking lines are corresponding to each case. The colouring of lines is according to the user-selected variable rule. By choosing one single line, such as that shown, the five variable values are displayed and the relation between variables is to be shown for the highlighted case.

Since Inselberg introduced PCP in 80's, much effort has been taken to enhance the capabilities (2009). Now the user can highlight filtering and colouring the variables they need through interactions. Discussion and implementation of the most used interactive tools can also be found in Edsall (2003), Andrienko and Andrienko (2004), Guo, et al. (2005) visualizing the multivariate patterns in a PCP display and a geographic map (GeoMap). In their work, it also suggests us human interactions to discover complex patterns in an effective way.

2.5. Point pattern analysis

Since the volunteer collected phenological observation data in from of spatial points, the spatial attributes are included in the dataset. If studied with the pattern of such kind of VGI data source, it is then related to point pattern analysis. Particularly when large volume data is presence, the complicated situation in its pattern is very likely appearing for VGI data source. Spatial point pattern analysis is a kind of study that is able to deal with complicated pattern cognition. In fact, the pattern analysis of spatial points interested researchers over decades; the theory in fact came out long ago.

Within a range of exploratory data analysis techniques and methods, point pattern analysis is one of an effective analysis, claimed by Smith et al. (2008). Here, the pattern of spatial point actually refers to the point form data, in which a point denotes the location (Delmelle, 2009). The term "point pattern" was initially defined by Hudson and Fowler (1966) as: *"the zero-dimensional characteristic of a set of points which describes these points in terms of the relative distances of one point to another." "Real world objects may be regarded as points if their areas are small in relation to their reciprocal distances."* For instance, particular occurrence of spatial phenomenon like crimes, diseases or fires used to be described in spatial point pattern.

The original technique utilized in the statistical analysis of point patterns was based on historical interest on plant ecology (Boots & Getis, 1988). Later, plant and animal ecologists used such techniques to explore both the spatial distribution and the interrelationships from more species. The main goal was just to identify individual factors and environment that could affect to the distribution pattern. It was also introduced to other areas such as archaeology, astronomy and anthropology in earlier 1980s.

Nowadays, the need of understanding point pattern spatiotemporally is gradually increasing. A concept of synchronization has been taken into consideration. Though synchronization is an important concept, the definition varies in the many fields, such as computer science, multimedia, photography physics and etc., For example in the social study, Wanstreeta & Steina's work (2011) suggests social presence is highly positively correlated to cognitive presence. In this sense, the synchronous discussion (presence over time in synchronous) for small-group, learner led discussion process is investigated. In this research, the related synchronization definition is the spatiotemporal synchronization of the species' phenological phenomena. This means for the observation illustrated in visualizations, there could be several synchronous or non-synchronous groups of point presence in space or time. Such study in phenological synchronization has not been applied on VGI data yet. This should pose the challenge to this research.

2.6. Apply GVA to VGI, issues to be concerned

To fully visualize and take advantage of VGI potentials with GVA techniques, a better understanding of how VGI characteristics impact the geovisualization is needed. This means, while the implementation, one should always take into consideration for VGI characteristics. Khalili et al. are interested in visualizing the geography of social networks (2009). The network is analysed from abundant VGI. They claimed that geovisualization has not been used to its full advantages yet because of VGI's challenging properties of being large and multivariate. Therefore, the method should be designed appropriately concerning these challenges.

In other fields, relationships can vary according to different opinions and viewpoints such as the emergence of conflict. The utility of VGI is starting to be much more used for crisis response and for decision making by governments. Again in this ecological aspect, growing concern has presence, e.g., the USA Phenology Network (NPN, 2011b) and the European Phenology Network (van Vliet et al., 2003). The challenges, of getting to know how opinions and viewpoints in the data can cause the effect to proceeding research, is concerned and regarded as an important issue.

To sum up, this research should explore to show mixed methods (GVA), mitigates each weakness and discover complex patterns in VGI datasets in an efficient way. Therefore, consistent thinking and concerning for VGI characteristics is ought to be taken.

3. DATA AND METHODS

3.1. Data

For exploring, visualizing VGI data, appropriate methods have to be considered. This chapter describes the data and the methods chosen to conduct the research.

3.1.1. VGI dataset

Two main types of datasets are used in this research, a VGI dataset of phenological phenomena and an environment dataset. The following subsections describe these datasets.

The dataset is called "Natuurkalender", for it comes from the Dutch phenological network with the same name (Wageningen UR & VARA, 2001). This phenological network is an environmental observation program that aims at studying the timing of the nature phenomena in relation to climate change. This network exists for 10 years (since 2000) and currently has about 8000 volunteers that collect phenological information all over the country. Volunteers record observations for species such as trees, herbs, insects and birds. It is recorded as phenophases (the visible stage for species life cycle development events). For instance, such phenophases within plants are including first leaf, budburst, first flower, last flower, seed dispersal, first ripe fruit, and leaf colour change. In this research, the recorded observation for 13 species kinds from Natuurkalender dataset is applied, including five birds, three butterflies, three herbs and two trees.

Species name (Dutch & English)	Phenophases(English & Dutch)	Category
Bonte vliegenvanger (Pied flycatcher)	First seen (Eerste individu)	Bird
Gierzwaluw (Swift)	First seen (Eerste individu)	Bird
Fitis (Willow warbler)	First heard (Voor het eerst gehoord)	Bird
Koekoek (Common cuckoo)	First heard (Voor het eerst gehoord)	Bird
Koolmeels (Great tit)	First young fly (Eerste jong uitgevlogen)	Bird
Citroenvlinder (Brimstone butterfly)	First seen (Eerste individu)	Butterfly
Dagpauwoog (Peacock butterfly)	First seen (Eerste individu)	Butterfly
Oranjetipje (Orange tip)	First seen (Eerste individu)	Butterfly
Bosanemoon (Anemone nemorosa)	First flower (Eerste bloei)	Plant (herb)
Speenkruid (Lesser celandine	First flower (Eerste bloei)	Plant (herb)
Fluitenkruid (Anthriscus sylvestris)	First flower (Eerste bloei)	Plant (herb)
Paardekastanje Witte (Aesculus hippocastanum)	First flower (Eerste bloei)	Plant (tree)
Eik zomer (Oak summer)	First leaf unfolding (Bladontplooiing)	Plant (tree)

Table 3-1. Subset 13 species in the dataset, ranging from birds to trees.

Table 3-1 shows the studied species in this research; ranging from common birds in Europe to herbs and trees. These species have recorded in different phenophases, ranging from first seen of species, first heard of species, first young fly away, first flower of the plant and first leaf unfolding of tree.

The geographical extent of this dataset is the whole Netherlands, which is also the studying area of the research. The coordinate system of dataset is Dutch National Coordinate system (RD-new). The dataset has temporal attribute like the date of observation; the temporal extent is from year 2003 to incomplete 2011 (simply because the research started before the end of 2011). Table 3-2 shows the dataset structure which has several attributes such as unique ID of observations (nr), species name (species), observed phenomena (phenophase), date of appearance (day), and name of appearance place (place), location of

appearance in x & y coordinates, scale of observation (the rounding level for x & y coordinates, indicating the last numeric coordinates are recorded in zero) and year number of observation.

nr	species	phenophase	day	place	x	y	scale	year
1660	Bonte vliegenvanger	Eerste individu	2006/4/6	Assen	233000	554000	1000	2006
1841	Bonte vliegenvanger	Eerste individu	2006/4/7	NP de Hoge Veluwe	185000	456000	1000	2006
1972	Bonte vliegenvanger	Eerste individu	2006/4/8	6525, Nijmegen	188000	425000	1000	2006
2028	Bonte vliegenvanger	Eerste individu	2006/4/10	7039, stokkum	212000	433000	1000	2006

Table 3-2. VGI dataset structure

3.1.2. Dutch environment dataset

As ecologists attempt to learn more about the environmental factors that phenophases respond to, in this research the study of how do plants and animals get to know the time of flowering, leafing and migrate have to be conducted to reach the third sub-objective. In this sense, several different types of environmental datasets are necessary for the research.

In the Netherlands, the Royal Netherlands Meteorological Institute (KNMI) delivers weather forecasts and warnings related to extreme weather events. In addition, KNMI collects data about climate and seismology and it conducts strategic and applied meteorological research (KNMI, 2011).

More precisely, the dataset consists of average daily temperature, sum of daily precipitation, sum of daily evaporation in the Netherlands. The selected KNMI dataset in this research is a gridded dataset (1km cell size) interpolated using daily data collected by about 150 meteorological stations throughout the years distributed all over the country Netherlands. These are critical influencing elements for the phenology study. The temperature is a measure of the warmth condition of the air. It is recorded at a measuring height of 1.50 meters. The average daily temperature is the mean temperature of 24 observations in a natural day. The maximum and the minimum of temperature are the highest and lowest of the measured values respectively. They are measured in degrees centigrade. Precipitation refers to the volume of rainfall reaching to the ground per square meter. As the amount of rainfall also considers the time duration (cumulative time), it is measured in unit kilogram per square meter per second (kg/m^2s). The evaporation (EV) or more precisely in this research the evapotranspiration (the evaporation from vegetation) is a vaporization that liquid water occurs on the surface of a liquid. It is measured in unit kilogram per square meter per second (kg/m^2s) (KNMI, 2004).

3.1.3. MODIS Land-cover dataset

As land-cover plays a vital role in climate and other fields of the earth system, the land cover has considerable control on the phenological cycle by significantly influencing the climate system through the reactive species (NASA 2011a). Further, variations in land cover generate variations of weather and climate by forcing atmospheric circulation patterns and the other way around, weather and climate can be influenced by the land cover. These patterns are driven by surface-atmosphere matter and energy fluxes and the momentum of the earth rotation.

The chosen MODIS Land cover product describes the land cover in various properties and is derived from observations ranging one year's input of Terra and Aqua satellites data. The MODIS product has 5 classifications and in one of them, 17 land cover classes are defined by the International Geosphere Biosphere Programme (IGBP) for the land cover scheme. The scheme includes three developed and land classes, 11 natural vegetation classes, and three non-vegetated land classes.

This research chooses MODIS combined satellites platform (Aqua & Terra) data Mcd12Q1.005 (last collection). The product has a raster structure with a cell size of 500 meters and its temporal granularity is yearly. The temporal extent from year 2003 to 2009 is selected for study.
In addition, the data is gained from Data Pool (NASA 2011b). The Data Pool is the publicly available portion of the Land processes distributed active archive centre (LP DAAC) online holdings. Data Pool provides a more direct way to access files by foregoing their retrieval from the near line tape storage devices. All Data Pool holdings are available at no cost.

3.2. Methods

In this section, the methods are presented one by one together with a brief introduction of their functioning and role in this thesis. The VGI characteristics are tackled through the combining techniques applications. The methods are arranged to answer the research question (see section 1.2.3). As figure 3-1 illustrated, the research work flow is composed of the following phases: data preparation, exploratory data analysis, self-organizing maps, multivariate analysis, VGI impact and human interaction interpretation).

1) Subset dataset annually into individual species and filter wrong observation for data preparation. Joining environmental variables with individual species observation in table files.
2) Obtain insight of VGI ground observation dataset along certain group of species. Annotate information to various species in each year, eg. number of observation or first observation day, last observation day.
3) Combine chosen geovisualization methods representing spatiotemporal pattern with basic statistical data analysis: space-time-cube visualization, geographic map and time-series map and basic statistical analysis techniques.
4) Generate self-organizing maps with prepared data. Identify the time synchronous group of species. Input time synchronous group in space-time-cube to detect synchronization in space perspective.
5) Analysis data joint with Dutch environmental data (temperature, evaporation and precipitation) on parallel coordinate plot for one species and annually interpret it with brushing. Find the annual space-environmental variable relations in three dimensional presentations.
6) Generalize the interpretation results and propose the impact from VGI.

The first step will create the filtered and subset table files for each species in annual year and a combined file for VGI observation and environmental variables in individual species. This helps the further method application part.
The second and third steps are to gain an overview of the data. Through this overview, one may get a general impression of the whole dataset and be able to identify the interesting part. The third step keeps the overview visualization, meanwhile focusing on the interesting pattern through time and space. This exploratory data analysis will identify and map spatiotemporal patterns in phenological dataset. And the first two research questions are to be answered in this exploratory data analysis.
For the analysis task laid in the research step fourth, the second sub-objective of research is corresponding to this step, apply SOM to the temporal profiles of each species for time synchronous group detection and identify spatiotemporally in space-time-cube.
The fifth step multivariate analysis performs an advanced exploration on the detail of the prepared data. It produces interactive visual techniques (brushing) and three dimensional presentations to combined data (VGI observation and environmental variables) to reinforce human cognition on finding out relations between species and environmental variables. This step enables to answer the last research question.

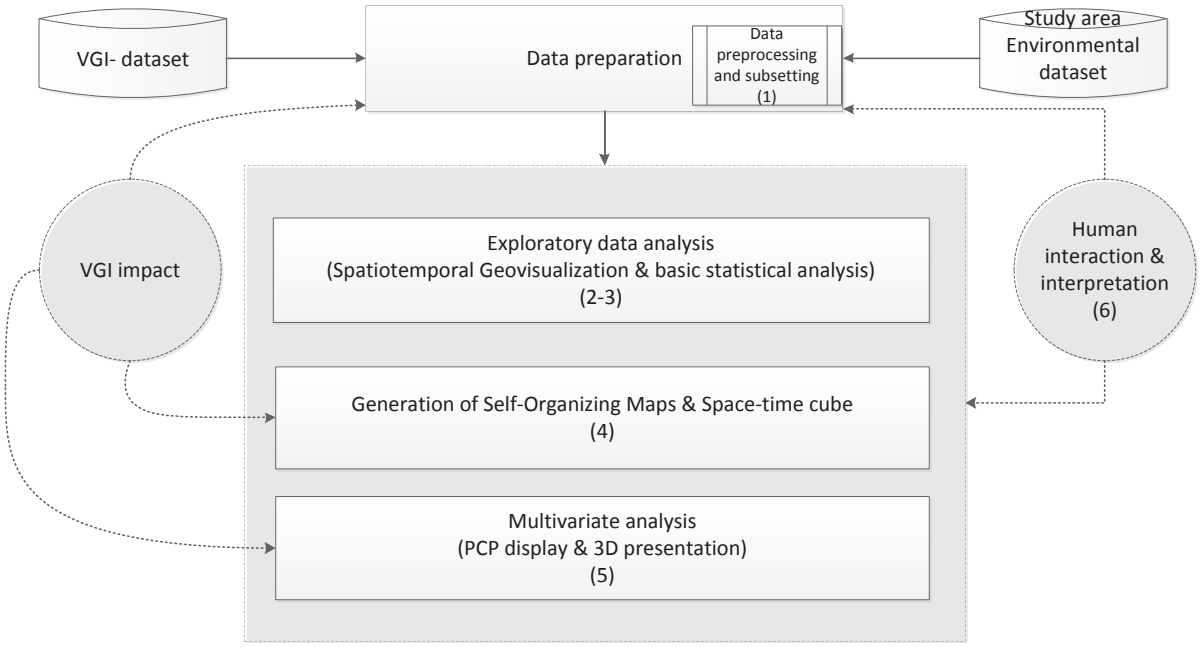


Figure 3-1. Methods work flow

3.2.1. Data preparation

The data preparation is always needed when using the regular data. As VGI has multiple characteristic (see section 2.2), data preparation is necessary to check the VGI dataset whether is ready for the coming analysis procedure or not.

Since erroneous observations can be contributed into the VGI data source (see section 2.2), the appearance of mistaken phenological records (incorrect observation) should be excluded in this research. The way is to check the spatial attributes x and y coordinates (the location of species). Thus, the filter is to be done and exclude the observation with zero value in x and y coordinates. The observation outside the studying area (the country Netherlands) is also to be filtered out of the dataset.

Since the most phenologists studied the phenophase in a recurring way (annually), a new temporal attribute has to be generated for the method better tackle the cyclical phenophase. Thus, for temporal attribute, the preparation will be done by calculating and adding the Julian-day (used in calendric calculation and present the interval of time in days and fractions of a day since January 1, 4713 BC Greenwich noon) attribute based on date of observation. In this way, the time attribute can present better sense of how exact observation appears in a cyclic year.

Since the preparation for answering our third research question, the information environment dataset and VGI dataset ought to be combined for the study of environmental relations. Before that, a selected threshold is done for cumulative temperature, cumulative daily temperature above zero centigrade and cumulative daily temperature above five centigrade (T0_sum and T5_sum). And it is selected to apply their cumulative sum of daily precipitation (precip_sum) and cumulative daily evaporation value (EV_sum) as well. This is mainly because the presence of phenophase is highly related to these cumulative variables (Gardliner, 2009). KNMI data was extracted for each day until the date of the phenological observation. The weather and land cover data is retrieved from original gridded data to each phenological observation. In this way that for each observations in VGI dataset, the attribute of temperature, land cover type, etc will be joined to obtain the combined dataset.

The table 3-3 illustrated how the environmental variables are joined with ID number of phenological observations.

nr	T0_sum	T5_sum	precip_sum	EV_sum	Land_cover(Rastervalue)
1660	245.485	131.7222	150.935547	55.64476	14
1841	288.3877	173.2682	187.244019	65.9522	5
1972	322.5065	196.4832	202.989548	71.06696	8
2028	320.8988	198.1654	204.006149	70.22451	14
2244	336.1745	207.9967	172.371475	77.32404	14

Table 3-3. Combined environmental data structure

The subset is to be done for performing further methods (see section 3.2). The subset on individual species annually is to be done. In this sense, the subset will have amounts of data indicating in table 3-14. The data preparation is completed and can proceed to the next step.

	Observation number of year												
	Bonte vliegenvanger	Gierzwaluw	Fitis	Koekoek	Koolmeels	Citroenvlinder	Dagpauwoog	Oranjetipje	Bosanemoon	Speenkruid	Fluitenkruid	Paardekastanje	Eik (zomer)
2003	25	132	120	117	31	289	185	161	67	160	73	43	23
2004	41	245	136	179	54	365	277	209	83	202	157	79	43
2005	49	265	198	231	75	313	190	205	105	279	178	87	58
2006	62	227	181	221	41	294	219	192	123	307	156	83	41
2007	56	302	178	267	130	366	290	262	117	303	157	86	59
2008	63	239	190	202	60	210	194	192	118	330	195	92	57
2009	45	202	164	197	52	290	211	253	109	259	148	73	50
2010	58	221	155	250	42	286	218	221	104	239	118	58	53

Table 3-4. 13 species observation subset. The table shows the observation number for each species in each year

3.2.2. Exploratory data analysis

At this step of work, methods are applied in exploratory data analysis (EDA). The two aspects from spatial to temporal are concluded, with both non-statistical representations (space-time cube, geographic map and time series map) and statistical ways.

3.2.2.1. Space-time-cube, geographic map and time series map

The EDA research begins with the non-statistical representation (space-time cube). The first goal is to map overall spatiotemporal pattern of observations, corresponding to first research question. Regarding to spatiotemporal attributed data, the means of representing is to plot each observation in a space-time-cube. At first, the subset data is able to be input. Three axes are to be defined. X, Y axis shall be representing its geographic extend (within Netherlands country boundary) and Z axis is representing the Julian day of observations. Thus, the spatiotemporal pattern can be visualized in this presentation.

Secondly, since the large volume of spatiotemporal pattern can be difficult to visually analysis in the cube, a series of from geographic map is used to gain spatial knowledge about various species in density patterns. Thus, in space perspective, the analyst can interpret the annual spatial pattern to specie.

For this geographic map, the goal is to learn and indicate the geographic high density area of observations and how the clustered observations are geographically located. This step a map calculated in kernel density of observations in studying area is generated. Kernel density calculates a magnitude per unit area from point features using a kernel function to fit a smoothly tapered surface to each point (ArcGIS, 2011). The working platform can be software Arcgis. In it, the kernel density function is from ArcToolbox - Spatial Analyst Tools - Density.

The example in figure 3-2 shows of generation geographic kernel density map for specie swift (Gierzwaluw). The subset of the data in this species is to be input through each individual year (from 2003 to 2009). And kernel density function is to be calculated with the following settings.
In following settings, the input layer, 1000m by 1000m cell size, unit meters, search radius 10km are to be set up. Larger values of the search radius parameter can produce a smoother, more generalized density raster than smaller values. While the smaller search radius parameter can produce a raster which shows more detail. Confirm and the mapping procedure will be then done. The result map shall indicate the geographic high density area of observations from year to year.

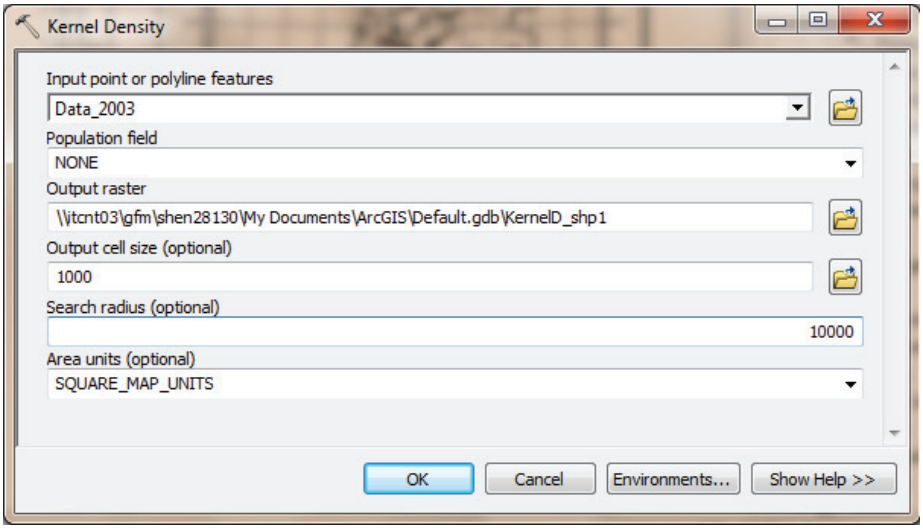


Figure 3-2. Function set-ups

Thirdly, time-series map is as a representation focusing on temporal perspective. Through it plot observations in two dimensions, the temporal pattern of species can be visualized annually. The meaning for time-series map is to identify overall pure temporal pattern of observations without distractions from space. In this way, the analyst can interpret. An example in figure 3-3 shows how it is implemented. In order to get map of specie, the subset of specie is input to analysis platform (R). Time-series map can be generated. According to Julian-day axis and ID number, time-series map can plot the observation in a temporal profile. With the brushing function, the map can allow analysts to do identification of attribute information for individual observation or groups of observations. The overall information of temporal pattern shall be identified in this way.

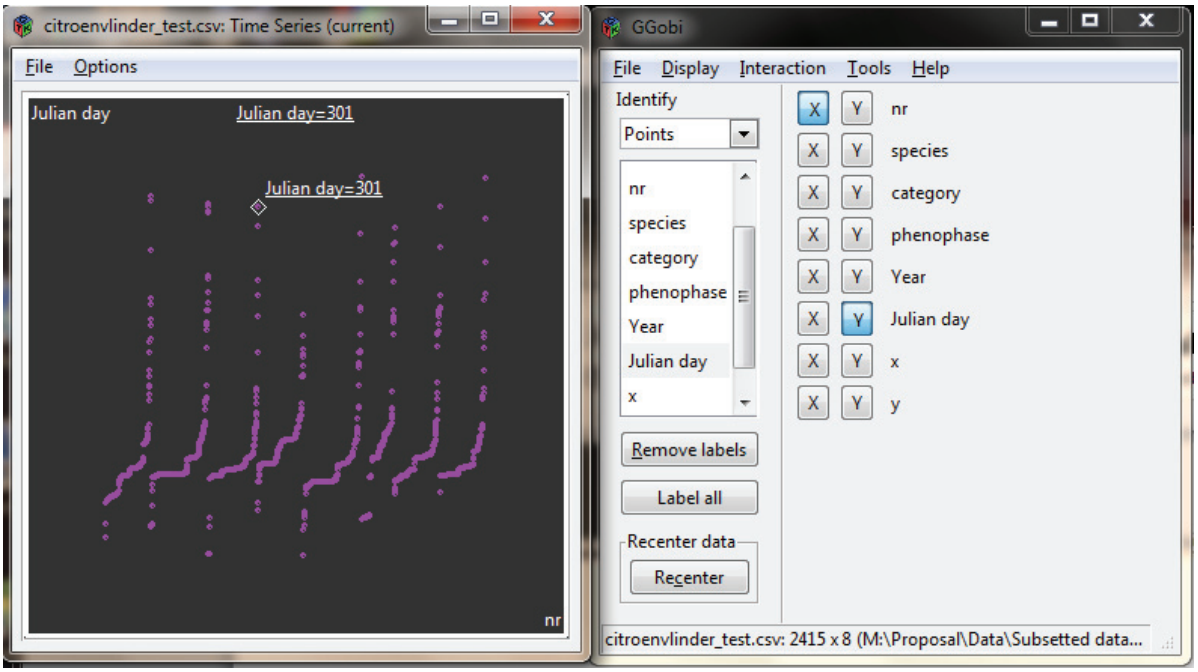


Figure 3-3. Time-series plot in R (ggobi package(Cook & Swayne, 2007))

3.2.2.2. Basic statistical analysis

In this subsection, a statistical aspect of data is to illustrate in boxplot and histogram for detecting temporal differences of species in statistics and causing effect from land cover type respectively.

Both of them are fairly common in descriptive statistics. A boxplot is a convenient graphical way of describing groups of numerical data. It has five levels of number summaries: the smallest observation (the minimum of sample), lower quartile (Q1), the median (Q2), upper quartile (Q3), and the largest observation (the sample maximum). A boxplot can also indicate outliers (observations out lied). The barplot plot the observation in sequence of time attribute (Julian-day). In this way, the observation from diverse time can plot as a temporal variation in year pattern for individual specie statistically. The other one, histogram, is a graphical representation showing an impression of the data distribution. With the prepared combined environmental variables dataset, the aim of it is to check the most land cover that observations was taken in. In this step of work, above two basic statistical methods are to be applied in analysis software R.

Firstly, for detecting temporal differences of species, the boxplot of observations in each year shall be applied. Different observations projected in this boxplot has Y axis for Julian-day of observations and X axis for year number. In this way, the variation of observation to different species can be visualized in sense of cyclic times (one year period).

Secondly, the histogram is applied to be visualized the causing effect from land cover type. The land cover type is already joined to dataset in data preparation part. Concerning the land type of Netherlands, water area should be excluded. This is because the VGI data in water area is insufficient and can mislead the results. Therefore no observations should be taken there. 16 land cover types are grouped from external MODIS Land-cover dataset into three basic groups. They are natural greenland for mainly various forest and grass land, semi human for area that are partially interacted with people, i.e., crop lands and the last one is urban built up where have the highest human interactions, i.e., city built up area.

In R, the barplot function can implement the histogram. The land cover type is recorded in raster value attribute. The histogram can show with Y axis representing the counting number of observations and X axis indicating the number of year. The year 2010 to 2011 data isn't taken into account as the incomplete data from MODIS Land-cover dataset.

3.2.3. Generation of SOM and STC

In this section, GVA techniques self-organizing maps and STC are chosen as the synchronization detecting techniques. For all species presented in the method, the time and space synchronization have to be explored. Thus, the second sub-objective is divided into two perspectives, time and space.

In the first part of analysis, the sub-objective is to study the temporal synchronization in phenological patterns of various species. A synchronous relation exists when observations occur at the same time. To study this, SOM is generated, a time-series one to group similar species in time. And then visualize the time synchronous species group in space-time-cube for the temporal synchrony confirmation and spatial synchrony detection.

Before applying the SOM to gain the temporal synchronous species, a time-series table is aggregated for constructing the temporal data pattern of all 13 species. The attribute Julian-day is used for generating its mean value and standard deviation. The two statics can represent the general temporal pattern of species through years.

species	Mean 2003	Mean 2004	……	Mean 2009	SD 2003	SD 2004	……	SD 2009
Bonte vilegenvanger	112.5	117.5	……	112.1	8.51	13.57	……	9.86
Gierzwaluw	116.4	116.4	……	120.5	4.81	4.92	……	9.83
Fitis	91.2	94.0	……	94.6	9.72	12.49	……	8.02
……	……	……	……		……	……	……	……

Table 3-5. Aggregated time-series data

As the table 3-5, the number of attributes is increasing in time perspective. In fact, the time domain from which these attributes originate and utilized is widening. Focused on time perspective of data, potential similarities are about to be revealed in SOM.

The Kohonen package from R is then used in the method. As this specific GVA technique, it can implement self-organizing maps together with some extensions for unsupervised pattern recognition. Since the result cannot be predicted, the unsupervised way is chosen for study. SOM approach can implement with large volumes, multi-dimensional datasets in an unsupervised way. According to Kohonen (2001) the SOM in its basic form produces a similarity graph of input data. Nonlinear statistical relationships are converted from multidimensional input data into simple geometric relationships with their image points on a low-dimensional display (reduction of dimensionality), usually displaying a regular two-dimensional grid of neurons. The neuron itself can learn and group the similar data according to attributes space. When the network is firstly setup with a grid, each neuron is randomly placed and has a vector known as codebook vector. The neuron in the self-organizing follows a kind of automatic machine learning procedure. This automatic machine learning procedure is in a way that first it chooses an output neuron that most closely matches the presented input vector, and then determines a neighbourhood of active neurons around the winner. The training of the SOM causes the codebook vectors to adopt the values of one of input vectors from data. The outcome codebook vectors of the SOM neurons are reshaped back to have characteristic data patterns. While some codebook vectors would not be the best matching unit for any input vectors. The empty neuron would indicate inter-vector distances. Eventually, it updates all the active neuron into a network. It is a neuron layer in which neurons are organizing themselves according to certain input values. This process iterates and it is the self-organizing. This learning procedure leads to a topologically ordered mapping for the input data. Similar patterns are mapped onto neighbouring regions on the map. While dissimilar patterns are located further apart.

Then the training of neural network can be conducted. To proceed with SOM analysis approach, there are several parameters that are critical to the final result, such as SOM grid size, the seed, the radius, learning rate, number of clusters and training iteration number,. The results drawn from the SOM mapping are fairly robust for different parameter settings (Wehrens & Buydens, 2007).

The first parameter, SOM grid, is the grid for the representatives. The size of it determines how many neuron lies in the SOM. As there are 13 kinds of species for training, the size is set up 3 by 4, which means there are 12 neurons for training. The second parameter seed is the random sampling seed. A certain value can keep neurons placed in a certain way. If set up in diverse value, the output of result can be slide different with the randomly placed neurons. The seed number is then set as 12. The radius is the radius of neuron's neighbourhood. When the neighbourhood gets smaller than one, the wining unit will be updated. In this research, default value (a value that covers 2/3 of all unit-to-unit distance) that set to radius is fit for research need. Another parameter learning rate is a vector of two numbers. It indicates the amount of changes. The default is to decrease linearly from 0.05 to 0.01 when training iterations take.
Another approach is used to determine the clustering number. This approach is one form of non-metric multidimensional scaling. It is called Sammon's non-linear mapping (Ewing & Cherry, 2001). It chooses a two-dimensional configuration for clustering groups. The input distance value is the distance of codes for our trained self-organizing-map. With this, the Sammon function can minimize the stress, the sum of squared differences between input distance and configuration. The configuration is eventually weighted by distances (RDocumentation, 2010).
The iteration number of training is another parameter should be determined in this SOM approach. It represents as the number of times the complete dataset is presented to the network. Therefore, in this research the network finally be mapped should have a well-trained and highly data correlated result.
The map of SOM needs a clear clustering line between neurons for a clear time synchrony interpretation. The approach is to apply the hieratical clustering function and set clustering line to group 12 neurons. Generally, no matter how much the network is trained, there will always be some difference between given input pattern and the neuron that is mapped to. The hieratical clustering function is a good way applied to investigate the sampling of data in training process. This clustering function groups similar neurons into hieratical structures. After the clustering number is determined, the hierarchical clustering can be finally used for adding SOM clustering boundary line. To cluster the codebook vectors, the method is, in R, calculate the hierarchical tree according to distance of SOM codes. Then, cut the implemented hierarchical clustering tree into initially determined number of clusters and reconstruct the upper part of the tree from the cluster centres. Give the neuron a clustering boundary with the former clustering tree. By this way, the neuron can be divided into groups. In these groups, species are sharing similarities in time.
The spatiotemporal representation in STC can then be implemented. This time, the time synchronous species group ought to be already known from the previous result. To confirm the temporal synchrony and discover the spatial synchrony within patterns, each subset species data will be plotted in STC by each year. The R package scatterplot3d is used for a visualization of data in a three dimensional space (space and time). The human visual interpretation in this step will be utilized for detecting the synchrony over space. A plane is extended for viewing the observation distribution both at space and time, aiming to get recognition for the annual species presence spatiotemporal trend. This plane is based on the linear regression, Julian-day over space (x and y coordinates) attributes, added in the cube, facilitating the interpretation. The appropriate annual interpretation and comparison between synchronous species pair can then obtain the spatiotemporal synchrony among species. In the meantime, the necessary discussion will be conducted on VGI characteristics in the result.

3.2.4. Multivariate analysis

The variability in daily temperature, precipitation and evaporation are often critical factors in phenological studies. These environmental factors may affect the timing of recurring species and lead a recurring pattern (see section 2.3). For the combined dataset (VGI observations with environmental variables, see section 3.2.1), a multivariate visualization will be applied in order to answer the third research question. The goal of this multivariate analysis is to find out the relationships between species pattern and environment variables.

The approach is divided into two parts to find out the relations to phenological spatiotemporal pattern respectively. First, using parallel coordinate plots (PCPs) the temporal pattern of individual species will be studied. With brushing, the annual pattern will be visualized. As trees and herbs are commonly sensitive to the environments, the case studied will be on a tree species paardekastanje whose flowering is normally believed to be highly influenced by cumulative temperature.

After studying the relationship between phenological temporal patterns and environmental variables, the relationship between space and environmental parameters is to be studied using 3D representations. In R, these representations are applied for the case study with same tree species aesculus hippocastanum (paardekastanje) and its environmental variables. This approach is to visualize the 3D cube and its linear regression plane, identifying the relation between space and three environmental parameters respectively. To sum up, through this mixed method, the third sub-objective can be fulfilled.

The PCPs are first to be applied as multidimensional data analysis. PCP display is used to gain a visual comparison between environmental variables in 2D annually. A normal PCP linearly scales individual axis representing each variable from the minimum to maximum. Other values are linearly positioned between the minimum and maximum on the axis.

The method is implemented in rggobi package. For analysis the multivariate attributes in a species, technique parallel coordinate plot can be applied to serves well for depiction of multidimensional data. To create such display here we can use the parallel coordinates display method on a ggobi data object. There has a need to specify the type of plot we want (the default is a XY Plot) and which variables to include. Since the joined dataset having environmental attributes like daily temperature, precipitation, etc., the display of overall pattern of all observation in PCP display can give analyst a view of interpreted relations (for example which variable is very critical that it has a relative narrow range). The attributes displayed in PCP are only four kinds and this would not lead too much difficulty for human vision to recognize patterns through many dimensions. But if the dataset is too large and data items overlap in the display, thus making the pattern hard to obtain, further interpretation of specific relationships needs interaction by the user.

The technique of bushing is recommended as interactive explorations while the interpretation. Brushing and PCPs are well composed graphic display techniques widely used to explore multivariate data (Inselberg, 2009). A brushing and identifying tool in rggob interface can help us highlight the observations of interest. For example, the high clustered plots in one variable temperature can be highlighted by brushing. In the meantime, these plots will show the relation on another axis such as evaporation. Therefore, the linking relationship between these two factors is displayed in one species per year. This can reduce the inherent visual cluster which is the main weaknesses in the PCP. In our human interpretation part, brushing will definitely play a vital role to detect multivariate patterns.

The second approach is to identify, for individual species in each year, the relation from environmental variable to space distribution of observations. The 3D visualization will be applied to compose the geographic space and attribute space (environmental attribute). Plotting annual observations in one species in the cube, shown in figure 3-4, three axes are representing the geographical attributes (x, y coordinates) and one environmental attribute respectively. These environmental attributes are sum of precipitation (Precip_sum), sum daily temperature over five degrees (T5_sum), and evaporation (EV). Therefore, three kinds of cube are displayed. And R package scatterplot3d will be applied in this step. The result will give a pattern of how trend appear between environmental variable and spatial observations. By comparing the annual pattern in the cube, the relationships between spatial observations and environmental elements will be revealed.

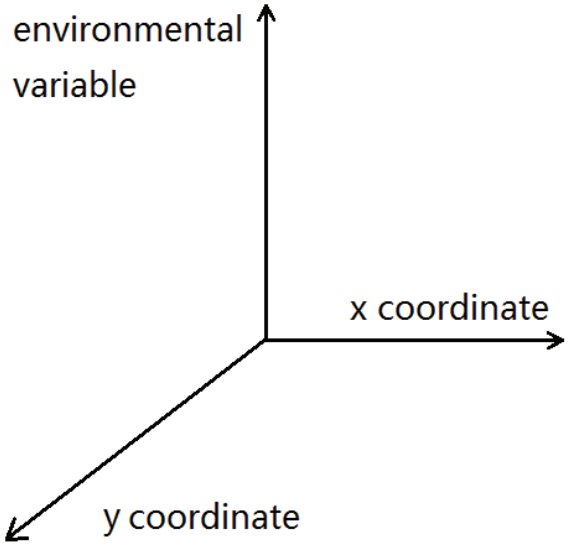


Figure 3-4. 3D presentation with geographical attributes and one environmental variable

4. RESULTS AND DISCUSSIONS

4.1. Introduction

This chapter describes the results coming from the analysis methods presented in chapter 3. The discussions corresponding to each step of the results are presented in subsections, ranging from exploratory data analysis and generation of SOM and multivariate visualization.

4.2. Exploratory data analysis

In this section, the dataset is described in figures from spatial to temporal aspects, with non-statistical representations (space-time cube, geographic map and time series map) and statistical ways. Recalling applied methods from section 3.2.2, the aim is to get an overall spatiotemporal understanding for pattern identification to the 13 species observations.

4.2.1. Phenological pattern in non-statistical presentations

The research firstly starts with non-statistical representations. The space-time-cube (STC) visualization for overall spatiotemporal pattern of observations is generated. The R package scatterplot3d helps to implement this representation.

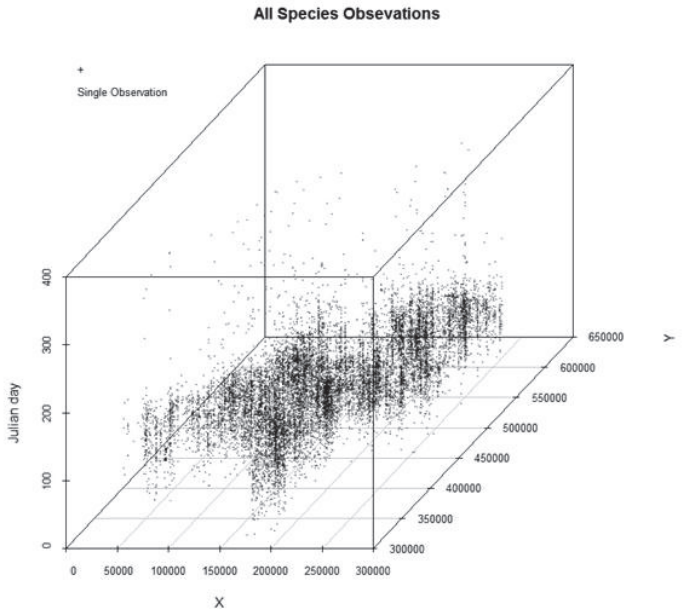


Figure 4-1. Space-time-cube visualization for 13 species observations (2003-2009)

The combination of 13 species observation through year 2003 to 2009 is presented in space-time-cube, in order to get an overview of phenologcial observations pattern in both space and time. There are three axes in this cube, x, y and the time axis Julian-day. The x and y axis together combined as the geographical plane. According to x, y coordinates and time attribute the space-time-cube is scattered plotted as figure 4-1. The scatterplot shows the relation for three variables. This result reflects the spatiotemporal presence of 13 species observations. As the figure shows, the plot of observation locates in the Netherlands geographically. The overall temporal pattern shows that the observations are clustered in a certain period of time. Generally, from spring to summer, the density of observations is the highest. This means that, during this time, volunteers are most motivated and interested by phenophases or that the species are

likely to be first observed during that period in the year. With the common sense in the region, the latter reason is expected. But it yet shows the interest of volunteers has impact to the pattern.
Secondly, to maintain the geographic information of observations, observations are naturally presented as dots in maps. And for the purpose to identify high density observation areas, maps are generated with the kernel density function.

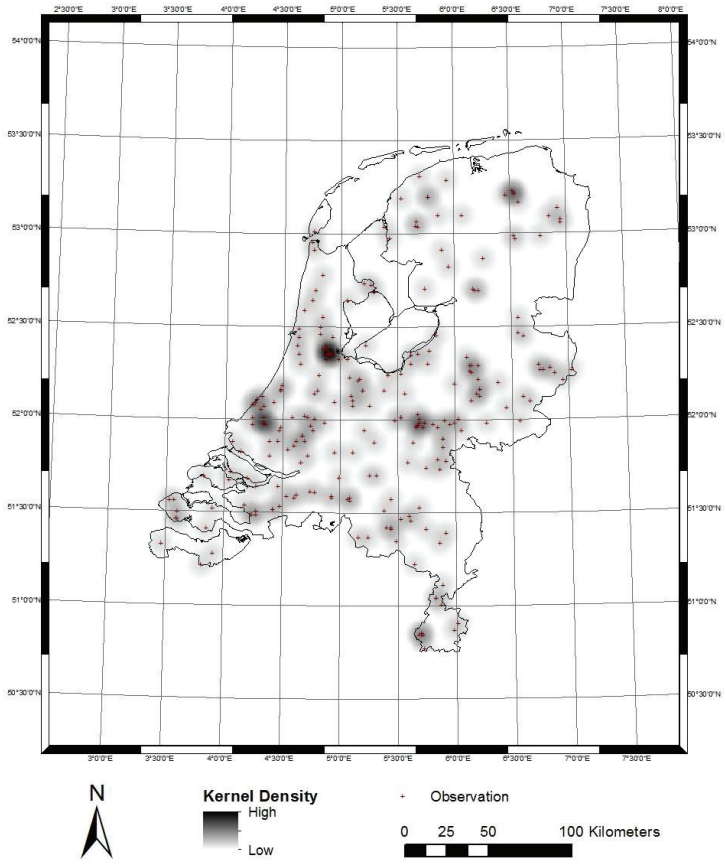


Figure 4-2. Geographical density map for swift (gierzwaluw) observations in 2007

The geographic map is generated for one species in each year. Take the result of bird species swift (gierzwaluw) observation in year 2007 as the example. The map shows not only the geographic distribution of each observation, but also the area with high density of these observations from the year. This helps analyst to understand the geographic pattern in a single year. With a map of Netherlands, analyst can identify that, the high density areas are located in following regions, Wageningen (in the centre of Netherlands), Groningen (north eastern part), Amsterdam (in the Northwest) and Maastricht (in the south). Among these cities, the region around Wageningen has the most frequent and high density in maps.
If compared with other years of observations in geographic map (Appendix A), a spatial pattern for individual species through years can be identified. This spatial pattern not only indicates the geographical distribution of observations. More importantly, I claim that it shows the VGI impact to the data. The above reigns are all cities in the Netherlands. And people in these cities act as volunteered phenological observers. The volunteers in the region of Wageningen show the most consistent interests in the Natuurkalender. Since this spatial pattern changes from each year, the presence of species is suspected to have elements affecting its changing distribution over times. These elements could be the changing

environmental variables, burst of interest from volunteers or migratory property of species like swift (gierzwaluw). This common bird spreads over the country following a migratory habit. The weather and food condition are the main reason for when the bird is begin to migrate. Meanwhile, since the species is common to be visualized during spring summer season, the repeated observation to same bird in the nest can be recorded. That depends on interest whether to take the observation or not. Therefore, for bird swift (gierzwaluw), both environmental variables and human interests can be great impact elements.
The studying area is the Netherlands. This means the variation in height is only within hundreds meters. The most places have relative even height. The height is often one critical element. As for the environmental variables like temperature is relative different in mountainous area which has huge difference in height, the phenological phenomena of species, particular for plants, are sensitive to the environment variation. However, with such small variation in the Netherlands, the height is not important issue in this study.

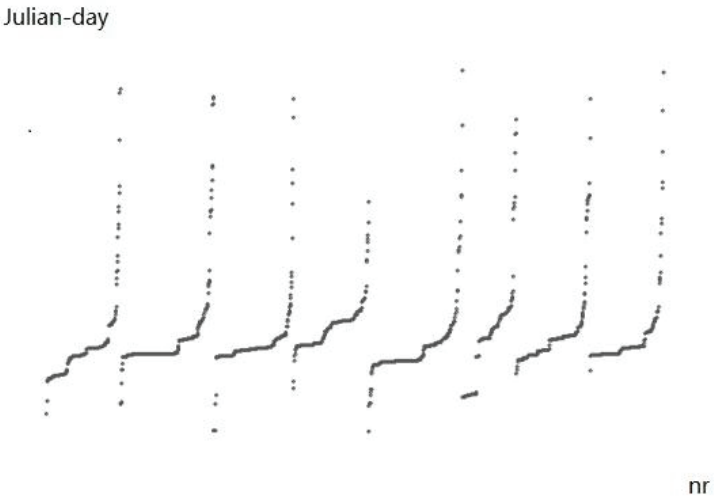


Figure 4-3. Time-series plot for Brimstone butterfly (Citroenvlinder) from 2003 to 2010 (left to right)

Thirdly, the time-series map gives temporal information of observations. The result is like a time profile for the species. Here we take one species from brimstone butterfly (citroenvlinder) as an instance.
As illustrated, figure 4-3 shows the temporal pattern from year 2003 to 2010. Among these years, the high clustering indeed appears in time of spring. As the vertical axis representing time variable (Julian-day) and horizontal axis representing ID number of observation, the lower slope of the profile indicates higher density of observations that lies in the time. This mean the species is likely to be first observed during that time of the year. However, slight variation can be identified in year 2008, in which a delayed time pattern shows up. The slope of this profile is larger than others. This change indicates the observation of species is not always stable. It can have variables influencing the time-series pattern. Thus, it is confirmed that there are one or several elements controlling the variation and affect the temporal pattern of observations. The assumption is both environmental and VGI impact could lead this yearly changes.
Through years, the observation dots are highly dense in the spring and summer. But individual observation points still presence in the very late dates. Since the observation for this butterfly is first seen by the volunteer, we assume the influence from wrongly observed by volunteers may lead to such situation. From this, we can assume that this proves the clustered of VGI data is much reliable for study.

4.2.2. Statistical analysis

In this subsection, results applied from basic statistical methods are reviewed. Firstly, a boxplot of observation is applied for species brimstone butterfly (citroenvlinder).
Two axes are shown in the boxplot. They are all time perspectives. As figure 4-4 shown, the mean value of Julian day lies in the range of time (75-100), which is the spring time in the Netherlands. For each year,

most of observations are indeed appearing in spring. But each year profile remains relatively unstable and is slight different from the other profiles. For instance, in 2008 observations have a larger range; this indicates a relative big variation in time. While in 2005, observations are highly clustered in a short period of days. This again gives us a hint that there are one or several elements controlling the variation of boxplot.

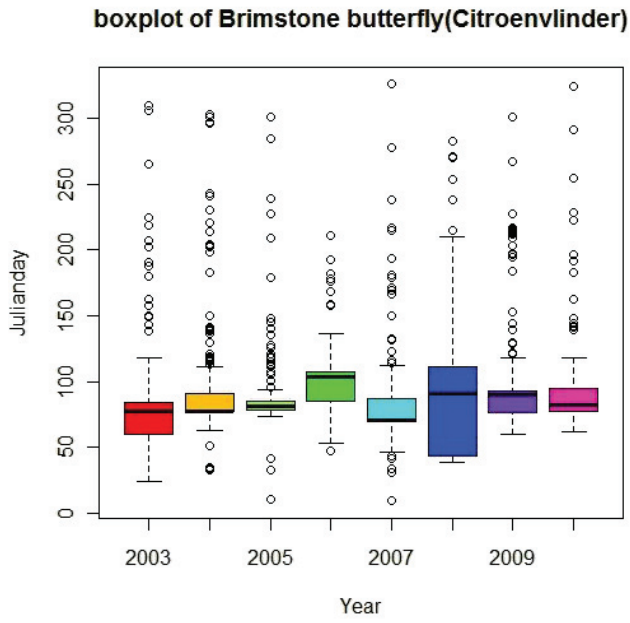


Figure 4-4. Boxplot for brimstone butterfly (citoenvlinder) observations (2003-2010)

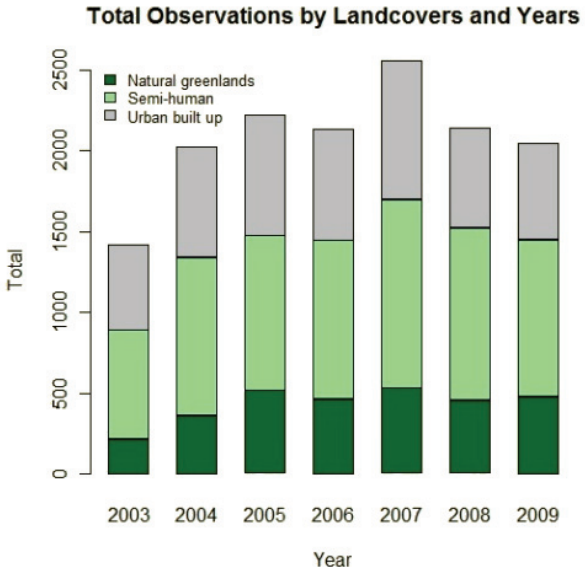


Figure 4-5. Histogram of land cover types for 13 species observation

Secondly, histogram is utilized to show how the land cover type is associated with observations. As stated in section 3.2.2, the result comes from the implementation with barplot function in R.

As figure 4-5 shown, two axes are year number and total number of all observations. In the year 2003, it has the minimum number of total observations and the maximum is in 2007. The number is generally

increasing from the first year until 2007. It then begins to drop slightly. The background knowledge for this trend is, the media like newspaper reported the Natuurkalender just before that year, and a burst of interest was taken place among people. The more people taking part in Natuurkalender, the more observations we obtained. Therefore, the VGI impact is influencing our results.

The other information is learned on land cover types. As the figure 4-5 shown, three aggregated land cover type are combined in the bar. Natural greenlands have the smallest proportion. The two other land cover types have human involvements in various degrees. Approximately 80% of the observations appear in these lands. This proves that human involvement is one of major elements influencing the distribution of pattern.

4.3. Synchronization in SOM and STC

In first part of this section, self-organizing maps is used for studying the relation among species in time. The relation is generated within an aggregated time-series table, in which the each species has its individual time variables through years (see section 3.2.3). The learning of which species show synchronization is studied in the following part.

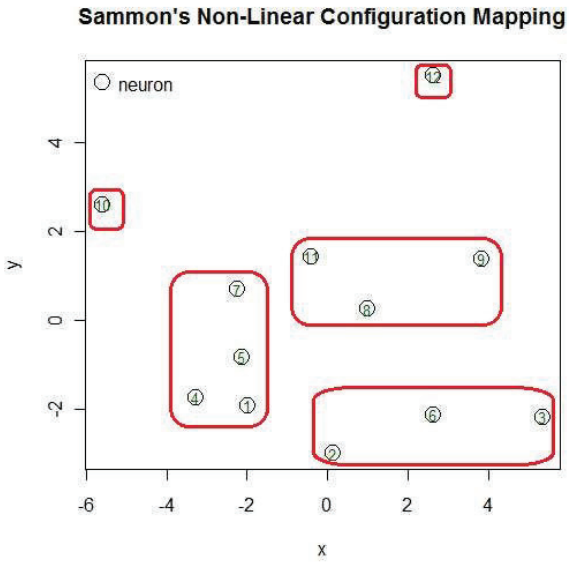


Figure 4-6. Sammon's non-linear mapping

Firstly, the aggregated time-series table is implemented in self-organizing map with the Kohonen package in R. The result of SOM is decided to be generated as 3 by 4 hexagonal grid maps (12 neurons represent 13 species data). In Figure 4-6, 12 neurons are distributed according to their mean distance of codes (see section 3.2.3 Sammon function). There are five clusters in the two-dimensional configuration. With human interpretation, a common learning can be obtained. Therefore the cluster number of species (see section 3.2.3 hieratical clustering) is setup as five for further SOM analysis. Five clusters are used to cluster the neurons, according to distance of codes in the SOM. With the determined parameter cluster number and other parameters setup in SOM mentioned in section 3.2.3, analyst can input the data in SOM network for training with one million iterations and get the following results.

Figure 4-7 shows various types of SOM plots for the trained network. The first neuron is the bottom right one and the second lays the left side of the same row. The fourth and seventh neuron is the left one in the second row and the right one in the third row respectively. The last neuron is the one at the corner top right of the network. Figure 4-7 (a) shows the sum of the distances to all immediate neighbours. This plot is also known as a U- matrix plot. The darker red colour the neuron gets, the lower average distances of

their neighbours. This implies how the neuron learns the time-series data. The black boundary line produced by heretical cluster approach (see section 3.2.3) divides the network into five different groups. Figure 4-7 (b) shows the number of species mapped to the individual neuron. The one with the grey colour is the empty neuron. It is the eleventh neuron and has no objects in it. And 4-7 (c) shows output vectors of each neuron. Figure 4-7 (d) shows the mean distance of objects mapped into a neuron to the codebook vector of that neuron. The red colour it gets, the smaller the distances, the better the neuron are represented by the codebook vectors. In this figure, we find that the fourth and fifth neuron have relative lower quality than others. The grey colour means no quality at all as the empty neuron shows. Figure 4-7 (e) shows, during the training iterations, how the mean distances to the closest codebook vector is declining. After one million iterations, the mean distance to closet unit declines to a certain level. Figure 4-7 (f) shows where the species are mapped and how 13 species are clustered.

With the results in figure 4-7, especially the figure (f), there are five clustering groups of species. From this, it can be identified that the five temporal similarities indeed exist among different species. Even though the species are from various categories (plants, butterflies and birds), some species share the temporal similarity and even synchronization.

For instance, in figure 4-7 (f) the lesser celandine (speenkruid, a herb), anemone nemorosa (bosanemoon, a herb) and the butterfly brimstone butterfly, (citroenvlinder) are grouped into one cluster. And first, sixth and seventh neuron are representing them respectively. In figure 4-7 (d), these three neurons have relative good quality and can represent the species data well. For these three neurons, the codes profiles are similar in figure 4-7 (c).

Another instance is that, in figure 4-7 (f), swift (gierzwaluw) and tree aesculus hippocastanum (paardekastanje) shares the same neuron. The two species swift (gierzwaluw) and tree aesculus hippocastanum (paardekastanje) are indicated a temporal synchrony. The two species are having synchronization through years. The same happens to bird species common cuckoo (koekoek) and pied flycatcher (bonte vliegenvanger). Therefore, it is known that the SOM trained these four species data in two neurons indicating much similar temporal pattern among them. But if compared to figure 4-7 (d), these two neurons, fourth and fifth, are the one with lower quality than others. It is learnt that the species data is not well trained in two neurons. Especially the fifth one, it has the lowest quality than others. In fact, the two species common cuckoo (koekoek) and pied flycatcher (bonte vliegenvanger) represented by it can be less clustered than the figure 4-7 (f) shows.

Species also have relative opposite characteristics in time. The great tit (koolmees) is the latest observed species and the lesser celandine (speenkruid) is the earliest one. The overall trend retrieved from the figure 4-7 is that, the species placed down right has the less mean of Julian-day for observations than upper left ones, which means the observations are earlier for downright species. The down left species orange tip (oranjetipje) has least variation on observed date (highly clustered in period of time) per year, while the upper right species peacock butterfly (dagpauwoog) has the most variation.

The profile in first neuron (down right) is opposite with the tenth neuron (upper left) in figure 4-7 (c). And in figure 4-7 (d), both neurons have relative good quality in representing species. This means the two species they individually represented, herb lesser celandine (speenkruid) and great tit (koolmees), have dissynchronization in their temporal pattern. The dissynchronization can be found in other two butterfly species orange tip (oranjetipje) and peacock butterfly (dagpauwoog), even though they both belongs to butterfly category. Therefore, this gives us learning that, no matter species are belonging to same category, species can share similarity and dissimilarity in time.

The result of self-organizing-map produced in R is robust when a series of parameters are setup as the same. These parameters include learning rate and grid size iteration time and more importantly the seed number (see section 3.2.3). The random sampling seed number has to be set as the same vale otherwise it will be very much influencing the layout of results, which may lead difficulty to get the exactly same result.

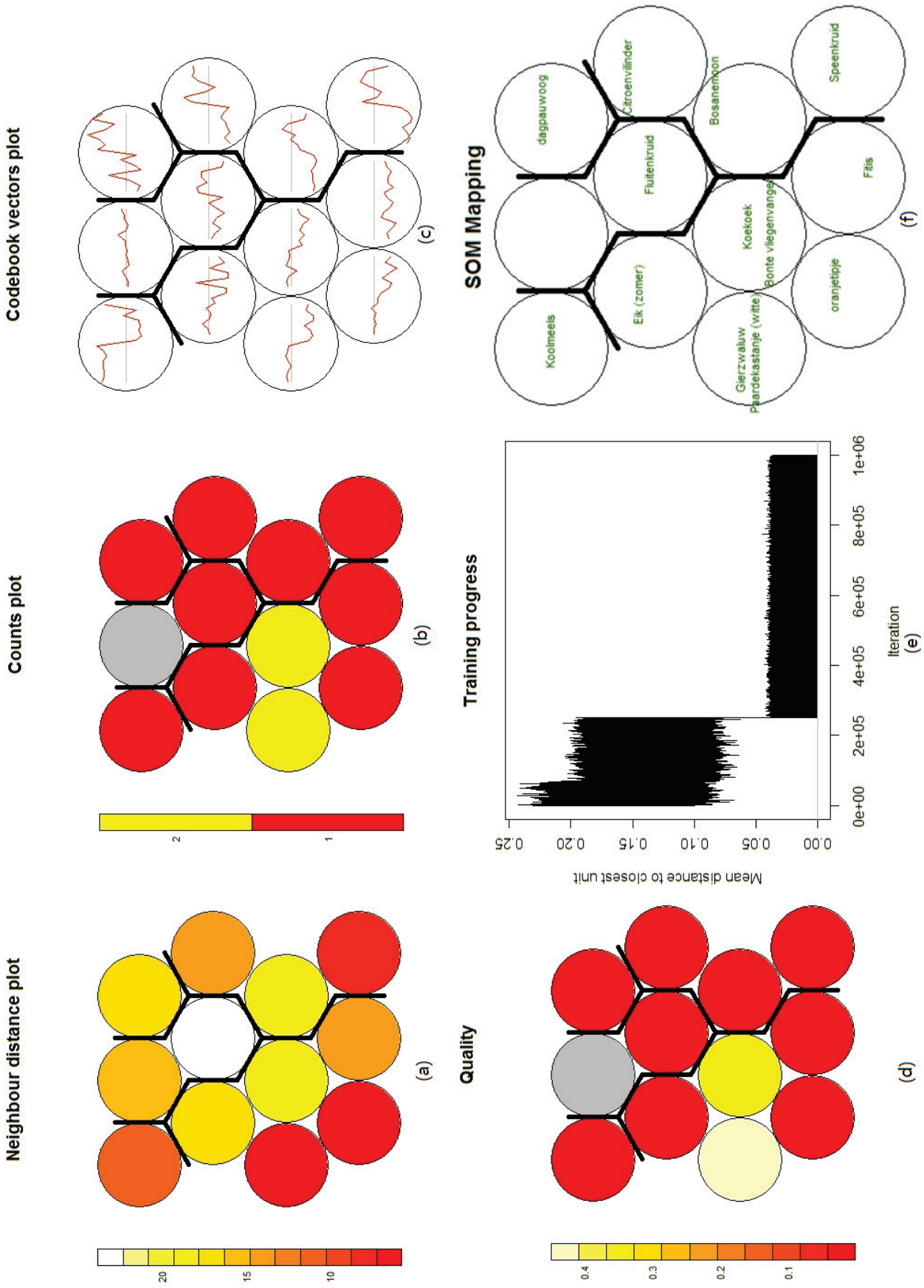


Figure 4-7. Self-organizing-maps

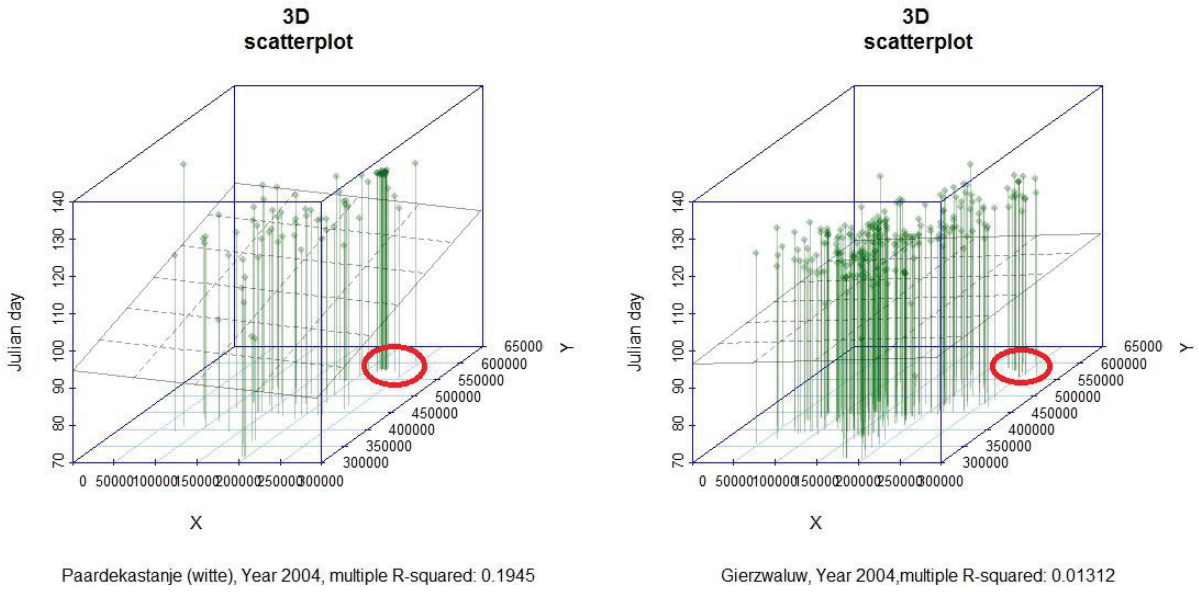


Figure 4-8. 3D scatterplot for species (tree aesculus hippocastanum-paardekastanje & swift-gierzwaluw) in year 2004 and a linear regression (Julian_day over x + y) plane for study overall spatiotemporal pattern

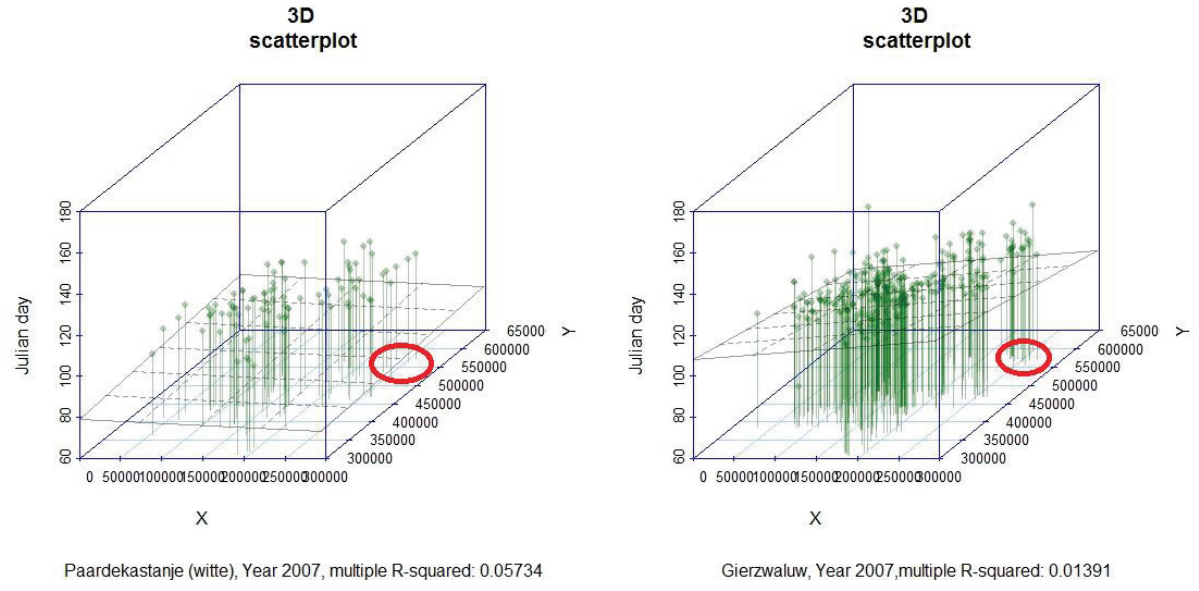


Figure 4-9. 3D scatterplot for species (tree aesculus hippocastanum-paardekastanje & swift-gierzwaluw) in year 2007 and a linear regression (Julian_day over x + y) plane for study overall spatiotemporal pattern

At this second step of research, analyst can obtain the temporal synchronous species from the SOM. Following annual spatiotemporal representation is to confirm the time synchrony and identify the space synchrony in STC for example species (swift and tree aesculus hippocastanum).

Figure 4-8, figure 4-9 shows two species observation distribution in 2004 and 2007 (Appendix A shows through seven years 2003-2009) spatiotemporally. The X and Y together constructed the geographic space of data. And the horizontal axis is indicating the time space (the attribute Julian-day of the observation). In this three dimensional visualization, both space and time perspective of view can be maintained. With the linear regression plane showing in the cube, the overall trend of observation can be visualized.

Through figure 4-8 and 4-9, the time synchronization through years between these two species can be confirmed. For instance, in 2004, for most of the presence of the tree flowering is observed just a few

days later than the bird. In 2008, they are sharing a rather similar range of time (around 120th day) to be observed.

Furthermore, the synchrony over space between species is discovered. For instance, in Figure 4-8, it is found bird swift and tree aesculus hippocastanum in 2004 shares one relative high space synchrony (co-location synchrony) in the north eastern part. And this space synchrony can also be clearly found in other year representations, shown in Figure 4-9. With the knowledge retrieved from the exploratory data analysis, the place is around the city Groningen (north eastern part of Netherlands). This can be because not only high interest of volunteers, but also the fact that bird swift indeed are likely to presence at the place, to avoid their predator and gain enough food they need. While the tree aesculus hippocastanum (paardekastanje) shows such a space synchrony in the cube as well. Especially in year 2004, at the same north eastern part, relative high space and time synchrony (both at around the 130th day of the year) is found. The presence of the tree phenomena is observed just a few days later than the bird.

Although not all regressions is well fitted with respect to multiple R squared value (mostly below 0.1), it is still important to facilitate the interpretation. For example, in year 2004, the tree aesculus hippocastanum (paardekastanje) has a temporal high corner in the north western part of regression plane. It implies the phenological phenomena of tree aesculus hippocastanum (flowering) firstly start from southeast to the northwest. Several observations are presented at very late time of the year (135th day+). The tree aesculus hippocastanum flowers earlier in the south-eastern than north-western in 2004. This pattern implies the controlling parameter is existed and we assume the environmental variables could lead such effect. This plane can help the interpreter find out the interesting space time trend for the observations. But the multiple R squared values are mostly below 0.1. It can require much more improvement like replacing it with a surface to better fit for most of observations.

For the migratory bird swift, the general migration pattern (in spring flying north to breed in the temperate or summer, and returning to the southern warmer regions in the autumn) is not entirely shown and cannot be retrieved from 3D representations. One reason is that, for our observation, it is recorded as first observed of phenomena. This means the representation can only show how migratory species is mainly firstly observed in spring or summer rather than observed their returning to south warmer parts in autumn. Therefore there is half of the bird general migration pattern in the presentation. The synchronization on time or place is studied based on this half migration pattern. The VGI also has impacted to the spatial distribution. Since the volunteer capture the information based on their interests, the place that has more interested volunteers implies contributes more influence to the overall spatiotemporal pattern we visualized. Thus, the plane shown in the cube not only shows the exact observation of phenological phenomena, but also concludes volunteers' variety interest in place.

Though this, the visualization successfully shows another pattern for swift (gierzwaluw). From the visualization, it is shown that most of observation lies around 120th day of the year. It matches the fact that, swift returns to its breeding place in their old colony approximately at the same time in each year, around the 1st of May (120th day in year) in central Europe (Driessens, 2010). The swift is known that it is very faithful to their breeding places. So the swift may breed in same place together in pairs for years. It is also matched the fact that repeated location of observation in different year that we can find within Netherlands. Such as area around the city Groningen, the presence of the swift species is relative commonly observed through years. The reason behind could be just simple that several volunteers in Groningen see the swift in a nest that is within the same house and repeatedly collect the information.

Though the comparison conducted with space-time-cube brings benefits for understanding space and time, in the meantime it includes some challenge. For instance, when we expect to compare the temporal synchrony between species with certain knowledge of space location, it is hard to visualize in two cubes in this case by interpretation. A possible alternative way is by presenting both species plots in one cube. But as the data size of VGI observation varies very much, the plot in one cube can be too massive and highly overlapped for interpretation. Just like this two species in Figure 4-9, in year 2007 the swift species has 302

observations and the tree had 86. To get equal size of data, random sampling to both species is not a very well option, as the method can decrease the variety and detailed information from the swift species side. The software R also has some limit for representation. The scatterplot3d packaged is not a well interactive package and need to be further developed. The results shown above are only static views of the STC. If one expects to interpret the STC with a personal understandable angle, the interaction like dragging the STC to a specific angle is not allowed. What is allowed in the package is only select a certain angle parameter and get the result after the coding. Therefore, it is not a user friendly package in this sense.

4.4. Multivariate analysis

In the last step of methods, the relation between species spatiotemporal pattern and variability in environmental variables like daily temperature, precipitation and evaporation is studied in the multivariate visualization. The approach is to find out the relations to phenological spatiotemporal pattern respectively. Firstly, temporal pattern of individual species is studied in parallel coordinates plots. With brushing, the annual pattern is able to be visualized. The case studied in the following is a tree species aesculus hippocastanum (paardekastanje) whose flowering phenomena are believed commonly sensitive to the environments.

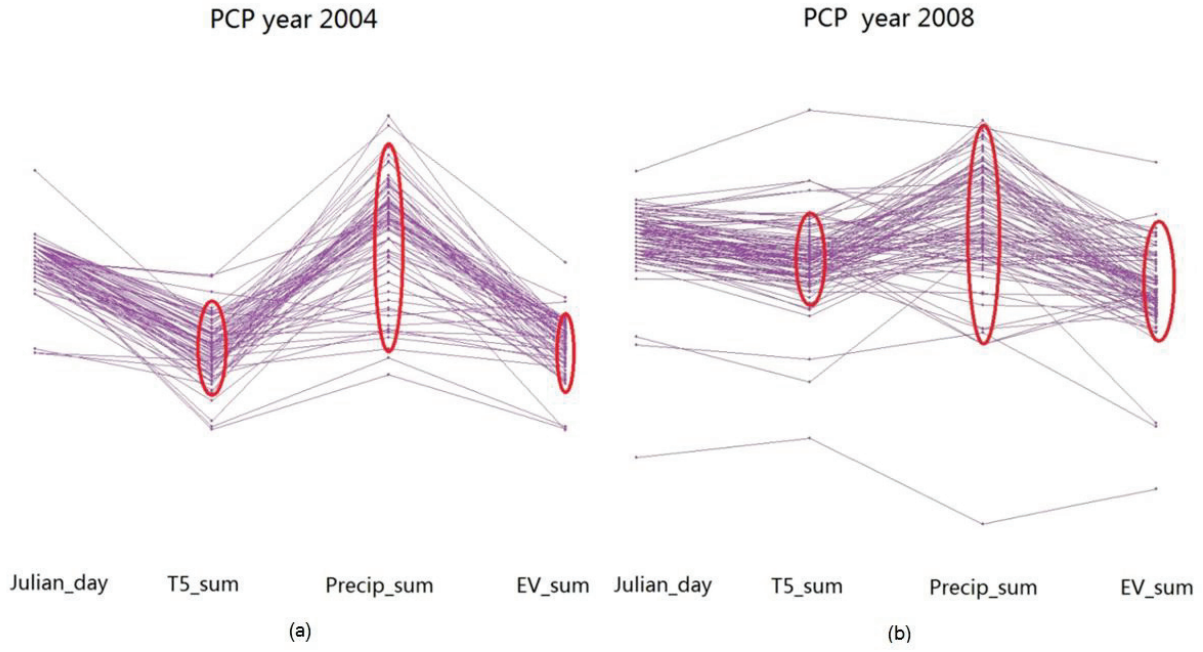


Figure 4-10. PCPs for tree aesculus hippocastanum (paardekastanje) in year 2004 and 2008 respectively (a) (b)

For instance, figure 4-10 displays the PCP for the tree aesculus hippocastanum's (paardekastanje) temporal attribute (Julian-day of the each observation) and connected environmental variables (above 5 degree cumulative temperature sum, precipitation sum and evaporation sum). The Y axis is represented by four variables in a range from a minimum to maximum (within 2003 to 2009) vertical scale and a purple linking line that corresponds to each observation. For example, in figure 4-10 (a), lines are coloured according to a chosen year 2004. Thus, in this PCP, other years' variables are shadowed and only present in 2004. From highlighted parts, shows a relation between the observation time and corresponding environmental variables. Thus, the aesculus hippocastanum's (paardekastanje) temporal pattern and its relation with three environmental variables are visualized. There are several out lied plots. They are apparently not well fit with the overall pattern through variables. But the meaning behind is the VGI data source lead such situation appear. The observation taken by volunteers with limit skills may not be correct. A very earlier flowering observation situation in this case can be that, a volunteer sees a flower that is not actually flower

for this tree but wrongly records it. But since the data quality is not an issue studied in this research, whether it is a false observation cannot be well justified. The VGI impact still exists in this part.
Another finding is that, through the indicated cycles, the range of temperature and evaporation values are relative narrowly connected with the species temporal attribute through years of PCP, from both two year plots in figure 4-10. It implies that, for these two variables, they can have very critical influence to the tree's flowering phenomena. As long as the temperature and evaporation reaches to certain levels, the flowering phenomena appear and are able to be observed. The narrow range finding confirms a common belief in ecology. The flowering of plant such as tree aesculus hippocastanum (paardekastanje) is very much sensitive to the temperature and evaporation.
After studying the relationship between phenological temporal pattern and environmental variables, the relationship between space and environmental parameters is studied in three dimensional representations. The three dimensional representation is applied for the case study with same tree species aesculus hippocastanum (paardekastanje) and its environmental variables. The approach is to visualize the three dimensional cube and its linear regression plane, identifying the relation between space and three environmental parameters respectively.
The example in figure 4-11 shows 3D scatterplot for tree species aesculus hippocastanum (paardekastanje) observations and their corresponding environmental variables (above 5 degree temperature sum, precipitation sum and evaporation sum) in year 2004 and 2008. The linear regression plane added for interpretation is calculated by environmental variable values over the spatial coordinate values. In temperature cube, the two year plane is rather stable. It indicates there are relative equal temperature sum values from place to place in Netherlands. The range in statistics is also almost same. Therefore, in a country level, the species aesculus hippocastanum (paardekastanje) has such kind of even relation in place through years studied. Similarly, the evaporation pattern in 2004 has almost no difference in 2008. The value range is almost the same. The linear regression plane in 2004 and 2008 are rather stable. This indicates a common even relation in evaporation value ranging from different places. The aesculus hippocastanum (paardekastanje) species needs almost the same evaporation for flowering in each part of Netherlands.
While for precipitation, the linear regression planes are relative more slant than the one in temperature cube. For precipitation, a clear trend is interpreted that the southeast part of Netherlands gets least precipitation value for presence of the tree flowering observation. As the space extends to northwest part, the precipitation needs to higher and it needs more rains to observe the tree flowering phenomena. This is a relevant learning to the geography of Netherlands. The south-eastern part is more humid than the north-western. The North Sea is besides the south-eastern part of Netherlands. In spring and summer season, the wind blows from south to the west bring humid air to the plant. The north-western land is less humid and needs to get more precipitation for plant to grow. Also, the two years precipitation range is very much similar. Therefore, the conclusion based on this is that the precipitation variable can control the presence of flowering observation in space. In spring summer season, the condition for firstly observing a aesculus hippocastanum flowering phenomena is less in precipitation value in the south eastern Netherlands than the north-western.
To sum up, the relation between temporal pattern of the species and environmental variables is obtained in PCP and discovery in three dimensional presentations is made, that tree species aesculus hippocastanum (paardekastanje) is very much sensitive to two variables (sum of above five degree temperature, sum of evaporation). The two variables critically control the flowering phenomena of aesculus hippocastanum (paardekastanje) evenly in space. The precipitation variable also controls the presence of flowering observation but differs in space.
The similar range of temperature and evaporation through years in Netherlands is led by some reasons. There are no huge mountains in Netherlands and the overall height varies only hundreds meters. The height of land plays one critical role in controlling environmental variables. With no such high variety in

height, the range of environmental variables can be similar small. Furthermore, the VGI influence still exists in space, recalling the discussion in section 4.3. One shortage appeared in this representation is that the static view limits our interpretation. A more interactive way for user to retrieve the information from representation is needed. However, the scatterplot function applied in R software is not highly developed. With only coding, it is recommended to develop a simple user interface for learning. Combining further animated view and interactive methods would be much more beneficial for this pattern comparison and recognition study

The size of data also can impact the interpretation for our results. The figure 4-11 shows the three dimensional observation patterns in country level. But one can be more interested in detecting the detail observation situation in certain location of Netherlands, such as a specific region. In this case, the visualization is failed. The reason is that the result cube cannot be zoomed in by an interactive way. The package just does not have that interactive function. The specific region which one expects to check at can even have no data collected. It also depends on the interests of volunteers and whether the observation records are collected or not. One solution might be combining other volunteered observation information to increase the size of data. But the variety of standards in data contributions from various data source can be one problem. All in all, the assumption for this can be the regional study of the phenological observation may depend on the interests of volunteers and the existence of the observation records.

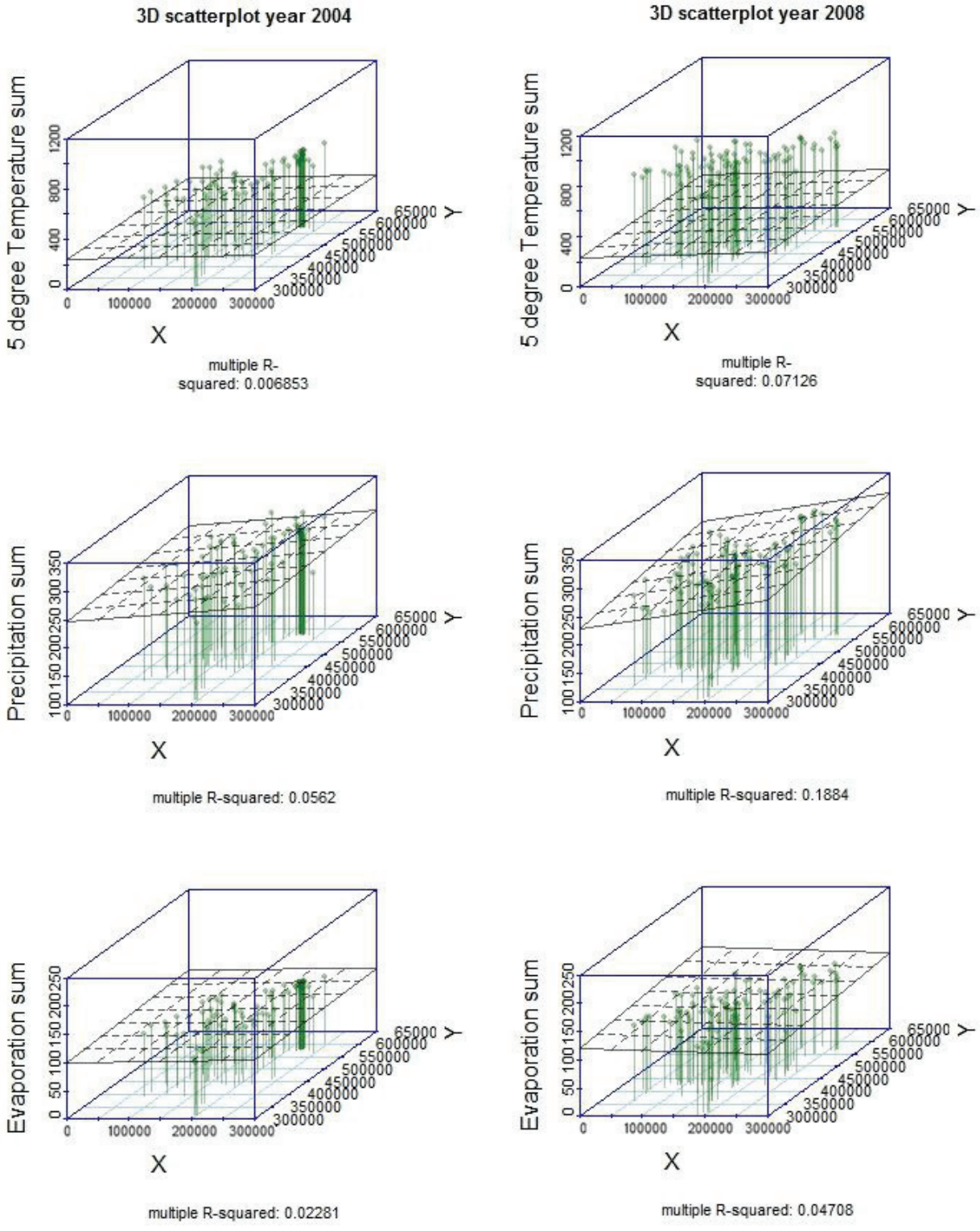


Figure 4-11. 3D scatterplot for aesculus hippocastanum (paardekastanje) with environmental variables in 2004 and 2008, added a linear regression plane (tempature_sum over x+y, precpipation_sum over x+y, evaporation_sum over x+y)

5. CONCLUSIONS AND RECOMMENDATION

5.1. Introduction

The main objective of this research was to find suitable geovisual analytic techniques for exploring spatiotemporal pattern in novel data source, phenological volunteered geographic information. This thesis explored a branch of visualization and analytic techniques on this spatiotemporal data source. Based on the main platforms R and ArcGIS, the applied techniques range from basic geographic maps, time-series plots to basic visualization environment as space-time-cube, clustering geovisual analytic technique as self-organizing-map, multivariate analysis as well as parallel coordinate plots. This chapter describes the conclusions of the thesis and the recommendation for future work.

5.2. Conclusion

In this research the main objective was achieved by fulfilling three research sub-objectives. Each sub-objective was achieved by answering following research questions.

First research sub-objective:

Identify and map spatiotemporal patterns present in the phenological dataset.

The related research questions are:

- How can we identify spatiotemporal point patterns of phenological phenomena?
- How can we visualize the phenological patterns of various species?

The idea was to identify and map the spatiotemporal pattern in (observation points) of phenological data through spatial clustering and temporal pattern comparison.

The exploratory data analysis with non-statistical presentations (basic geographic maps, time-series plots and basic space-time-cube visualization) and statistical analysis (boxplot for temporal distribution of observation and histogram for observation taken in diverse land cover type) was applied to visualize the overall spatiotemporal pattern from the species under study. The integrated geographic knowledge discovery in exploratory data analysis was able to identify the spatiotemporal point patterns. Meanwhile, statistical analysis such as boxplot backup the discussion of individual species pattern.

The overall temporal pattern for the volunteered observations were mainly clustered in a certain period of time (spring summer season). The spatial pattern through years, such as bird species Gierzwaluw (swift) observations, indicated the high density around following regions, Wageningen (in the center of Netherlands), Groningen (northeastern part), Amsterdam (in the Northwest) and Maastricht (in the south). Among these cities, the region around Wageningen has the most frequent and high density in maps.

To sum up, the methods used in this step focused on different perspective of phenological pattern. The result from geographic maps can only present the spatial pattern in the geographic space. The time series maps can only tell the temporal pattern but do not have a statistical view of temporal variation through years. The space-time-cube can tell spatial temporal pattern but encountered with difficulty on interpretation and revelation for very large volume of data. The statistical analysis can bring the temporal variation per species per year in a statistical way for support. These methods cannot work well on visualizing and identifying spatiotemporal patterns present in the phenological dataset individually, but have to be combined to mitigate each other's weakness.

In a sufficient way, the spatiotemporal pattern in volunteered geographic information was identified and overall phenological pattern in space and time were visualized.

Second research sub-objective:

Study the synchronization of phenological patterns of various species.

Related research question is:

- How can we compare (e.g. find synchronous groups) phenological patterns?

The idea of the approach was to study the synchronization in pheonological patterns of various species. The same species pair different years were used for temporal comparison to get temporal synchronous pair. In spatial temporal patterns, pairwise comparison in space and time was done to species group.
In my approach, the applied techniques were self-organizing-maps (SOM) and space-time-cube (STC). The self-organizing-maps focused on detecting and clustering temporal synchronous species through years (2003-2009). The 13 species' similarity in time was revealed by input trained aggregated time-series data to SOM. Five groups of species sharing with various temporal synchronies were produced. Temporal synchronous species groups are found such as group one-the lesser celandine (speenkruid, a herb), anemone nemorosa (bosanemoon, a herb) and the brimstone butterfly (citroenvlinder); group two swift bird (gierzwaluw) and tree aesculus hippocastanum (paardekastanje). The dissynchronous species groups in time were also found, like group one-herb lesser celandine (speenkruid) and great tit (koolmees, bird); group two- two butterfly species orange tip (oranjetipje) and peacock butterfly (dagpauwoog).
After that, the STC was used to confirm the temporal synchronization and detect spatial synchronization between temporal synchronous species pair (swift bird (gierzwaluw) and tree aesculus hippocastanum (paardekastanje)) through interpretations. Through annual comparison in STC, this temporal synchronous species pair was confirmed to have time synchronization within years (like in 2004, for most of the presence of the tree flowering is observed just a few days later than the bird on 120th day of the year). The space synchronization lied in places such as Groningen (north eastern part of the Netherlands).
The two methods used in this step focused on different perspectives. The SOM was a powerful clustering methods in attribute space and able to cluster the temporal synchronous species. The space-time-cube cannot simply groups the synchronous species, but can link data from the geographic space to time. With a support of linear regression plane, the analyst can detect the synchronization in space between species pair (swift bird (gierzwaluw) and tree aesculus hippocastanum (paardekastanje)) with visual comparisons. The two methods both had weakness and strengths. The approach was to integrated them and make them focus on their strengths part. In this way, synchronization of phenological patterns was studied and second sub-objective was able to be achieved.

Third research sub-objective:
Discover relationships between phenological patterns and environmental factors.
Related research question is:

- How can we represent the relationship from environmental factors to the phenological phenomena pattern?

In the last step of research, the multivariate analysis was applied and represented the relationship from environmental factors to the phenological pattern of tree aesculus hippocastanum (paardekastanje) flowering observations. Three environmental variables (sum of above five degree temperature, sum of evaporation and sum of precipitation) were studied in the approach. With the parallel coordinates plot (PCP), three environmental variables were linked to the temporal attributes of species. The relation between temporal pattern of the species and environmental variables was obtained and discovery of plant species is very much sensitive to two variables (sum of above five degree temperature, sum of evaporation) is made. Facilitated with multi-dimensional visualization in three dimensional cubes and liner regression planes (environmental variables over space attributes), the space attributes of species were linked with environmental variables annually. The general relationship for pattern of species and environmental variables were obtained. The narrow ranges in the parallel coordinate plots indicated that two variables (cumulative sum daily temperature over five degrees (T5_sum), and evaporation (EV)), critically controlled the flowering phenomena of paardekastanje (aesculus hippocastanum). The precipitation variable (sum of precipitation (Precip_sum)) also controlled the presence of flowering observation but differed in space. It concluded, for tree species paardekastanje, the temperature and evaporation evenly affected the presence

of flowering phenomena within the Netherlands. The precipitation for the tree flowering was different from southeast to northwest of the Netherlands.
The statistical analysis, histogram for the observation in diverse land cover type, in exploratory data analysis showed the interests of volunteers. The interests were shown in regions, Wageningen (in the center of Netherlands), Groningen (north eastern part), Amsterdam (in the Northwest) and Maastricht (in the south) and approximately 80% of the observations appeared in lands that has human interactions in various degrees. This concluded that human involvement is one of major elements influencing the distribution of phenological pattern.
The methods used in this step focused on different perspectives. The PCP can be visualized and interpreted on its showing pattern. With interactive brushing technique, the relations among presented variable was able to be interpreted in attribute space of data. While for 3D presentation, the individual environmental variable was linked with geographic space to discover relation to the phenological pattern. They cannot do all the tasks all by themselves. The best way was to mix the methods and the goal of this step was achieved.
The general conclusion is that, this research shows that "mixed" methods can mitigate each other's weakness, discover complex spatiotemporal observation pattern and explore the volunteered phenological datasets. For phenologist, this research expands a horizon for phenological phenomena pattern analyst by integrating visual, computational and cartographic methods together to detect and visualize spatiotemporal phenological observation pattern. For researchers in geographic information science, this exploration tells VGI potentials can be revealed.

5.3. Advantages and disadvantages

Advantages in the research lie as follows. 1) Exploratory analysis provides the necessary overall learning for further interpretational analysis. 2) Annual kernel density maps provide clear VGI impact to the overall spatial pattern of species. Particular for high density area, main cities can be detected and compared in different year. 3) Addition information can be retrieved from land cover histogram, giving a backup knowledge of large VGI influence lies in the phenological spatiotemporal pattern. 4) In an unsupervised way, Self-organizing-maps detect temporal synchronous species very effectively. Information like dis-synchronous species groups is detected easily. 5) The additional linear regression plane in three dimensional representation support the interpretation overall pattern recognition. This also makes data detection become easier with large volume of data in the representation. 6) Multivariate relation for a species observation in parallel coordinates plots is easy to be visualized through years with brushing technique. 7) The VGI characteristics such as large volume of data, spatiotemporal, point data form and interests from volunteers' opinions are included with appropriate methods. 8) The results from space-time-cube are not only shown the actual annual spatiotemporal pattern for species, but also the volunteers' interests on the data collection. 9) R software is rather robust for producing the exact visualization results with a series of parameter setup in coding.

The disadvantages also appear in this research. 1) The data quality issue is not included in this research. With no reference data, the quality of data source is unjustified. This leads the difficulty to identify some false observations. The bias of observation is not included. 2) The VGI dataset has 18000 observations in total. But for each species the number of observations differs through years. The issue on size of data is not studied in this research 3) Though this research proves SOM is a powerful technique to retrieve clustering data, there was only one self-organizing-map applied. 4) Although, linear regression plane plays an important role to facilitate the interpretation, not all regressions is well fitted with observations respect to multiple R squared value (mostly below 0.1) and it can requires much more improvement. 5) The result of STC is only static view and some further interpretation in detail location can be difficult to analyst. For example, the change of viewing angle needs to be done only in coding with R. Much more helpful

animated STC can be developed. 6) The effect of height is not included. For other study area, this can be challenging. 7) R software requires similar coding experience and no user interface is suitable so far.

5.4. Recommendation

After the realization of the research, certain techniques and objectives were not enough to complete. The flowering list presents the explanations of them as well as additional ideas which can be fairly critical to this research.

- **Bias of observation:** The bias of observation is considered to be one important issue that affects research result in some extent, though it is not in one sub-objective. Regarding to limit of time, the issue is not allowed to be investigated. On the other hand, in the case studied (spatiotemporal representation), the spatiotemporal pattern are affected by the bias of VGI data. Though the bias is within kilometres, this characteristic still exists and is essential for the provincial or regional level research. This implies a deeper field into the VGI data itself and deviate another direction of the study (VGI pattern research in regional area).
- **Size of data:** The VGI data source here applied in the research is only from Natuurkalender (Wageningen UR & VARA, 2001). This limits the size of species data we can applied in methods. Challenges on different size of species data can appear among the analysis. The difficulty for comparison can be encountered in interpretation. The solution seems to simply increase the size of species by adding other VGI data from various sources together. So far for Netherlands, the known possible VGI data sources collected based on phenological phenomena are National databank (Telmee, 2007) and Waarneming (Vries & Verheul, 2006). This implies in regional VGI data study the size of data can be sampled to fairly enough equal size for analysis.
- **Other self-organizing-maps:** There are other types of SOM. For further study, different types of SOM such as superSOM (Wehrens & Buydens, 2007) should be explored and studied in results comparison. For example, efficiency on very large dataset's result production can be check on SOM applications to select the most promising SOM for studying very large volume VGI.
- **Animated representation:** The representations in this research are all shown in static ways. This limits the human interactions especially for multidimensional representation results. In the future study, the representation shall be much more animated and interactive for visualization. For instance, it can be implemented on platform software R, analyst develops a friendly user interface that can allow user interact with the visualization results. With zoom in zoom out, drag function, these simple interactions can very well benefit the human interpretation. There is also other solution that is finding other appropriate software to implement the research goal.
- **Plane or surface, a support for multi-dimensional presentation:** In this research, one interesting attempt is the combination with a linear regression plane in the STC facilitated for overall pattern study. The relation between several dimensions can then be apparently visualized. But the fact is such regression plane cannot well fit with all observations. It can only indicate an overall trend. If detailed relation information in specific place or time is demanded, such plane is failed. The inspiring idea is that, instead of combine additional plane, a surface can be generated in STC for observation point pattern study. This will bring more valuable learning of how dynamic change in spatiotemporal pattern takes place from both time and space. With the idea proposed and the techniques available in this research, the only difficulty seems to lie in finding a way to implement in R codes.
- **Height:** Since our studying area is the Netherlands, the variation in height is only within hundreds meters. However, height element can be a challenging issue for studying data collected in mountainous area. The researcher needs to be aware the issue is very essential to study in high variation studying area. As for the environmental variables like temperature and precipitation is relative different in area with huge difference height, the phenological phenomena of species,

particular for plants, are very much sensitive to this environment variation. Their spatial temporal pattern can be fairly diverse in such special area.

LIST OF REFERENCES

Andresen, M. A. (2009). Testing for similarity in area-based spatial patterns: A nonparametric Monte Carlo approach. *Applied Geography, 29*(3), 333-345.

Andresen, M. A., & Malleson, N. (2011). Testing the Stability of Crime Patterns: Implications for Theory and Policy. *Journal of Research in Crime and Delinquency, 48*(1), 58-82.

Andrienko, G., & Andrienko, N. (2004). *Parallel Coordinates for Exploring Properties of Subsets.* Paper presented at the Proceedings of the Second International Conference on Coordinated \& Multiple Views in Exploratory Visualization.

Andrienko, G., Andrienko, N., Bremm, S., Schreck, T., von Landesberger, T., Bak, P., & Keim, D. (2010). Space-in-Time and Time-in-Space Self-Organizing Maps for Exploring Spatiotemporal Patterns. *Computer Graphics Forum, 29*(3), 913-922.

Andrienko, G., Andrienko, N., Dykes, J., Fabrikant, S. I., & Wachowicz, M. (2008). Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research introduction. *Information Visualization, 7*(3-4), 173-180.

Andrienko, N., Andrienko, G., & Gatalsky, P. (2005). Chapter 10 - Impact of Data and Task Characteristics on Design of Spatio-Temporal Data Visualization Tools. In D. Jason, M. M. Alan, A. M. M. Menno-Jan KraakA2 - Jason Dykes & K. Menno-Jan (Eds.), *Exploring Geovisualization* (pp. 201-222). Oxford: Elsevier.

Ankerst, M., Berchtold, S., & Keim, D. A. (1998). *Similarity clustering of dimensions for an enhanced visualization of multidimensional data.* Los Alamitos: Ieee Computer Soc.

Anselin, L., & Rey, S. J. (2010). Perspectives on Spatial Data Analysis. In L. Anselin & S. J. Rey (Eds.), (pp. 1-20): Springer Berlin Heidelberg.

ArcGIS. (2011). Kernel Density (Spatial Analyst) Retrieved 2011, December 2, from http://help.arcgis.com/en/arcgisdesktop/10.0/help/index.html#//009z0000000s000000.htm

Batty, M., & Longley, P. A. (2003). Researching the Future of GIScience. In P. A. Longley & M. Batty (Eds.), *Advanced Spatial Analysis: the CASA Book of GIS* (pp. 427-435). Redlands, California: ESRI Press.

Bjørnstad, O. N., Ims, R. A., & Lambin, X. (1999). Spatial population dynamics: analyzing patterns and processes of population synchrony. *Trends in Ecology & Evolution, 14*(11), 427-432.

Boots, B. N., & Getis, A. (1988). *Point pattern analysis* (Vol. 8). Newbury Park etc.: Sage.

Brian M. Tomaszewski, Anthony C. Robinson, Chris Weaver, Michael Stryker , & Maceachren, A. M. (2007). *Geovisual Analytics and Crisis Management.* Paper presented at the Proceedings of the 4th International ISCRAM Conference, May 13-16, 2007.

Cook, D., & Swayne, D. F. (2007). *Interactive and Dynamic Graphics for Data Analysis:With R and GGobi*: Springer-Verlag.

Delaney, D. G., Sperling, C. D., Adams, C. S., & Leung, B. (2008). Marine invasive species: Validation of citizen science and implications for national monitoring networks. *Biological Invasions, 10*(GEOBASE), 117-128.

Delmelle, E. (2009). Point pattern analysis. *In: International encyclopedia of human geography. / ed. by. R. Kitchen and N. Thrift. Amsterdam, Elsevier, 2009. ISBN 978-0-08-044910-4 pp. 204-211.*

Deparday, V. (2010). Enhancing Volunteered Geographical Information (VGI) Visualization with Open Source Web-Based Software Retrieved 2011, October 8, from http://hdl.handle.net/10012/5709

Driessens, G. (2010). The Common Swift Retrieved 2011, December 25, from http://www.commonswift.org/swift_english.html

Dykes, J., MacEachren, A. M., & Kraak, M. J. (2005). *Exploring geovisualization.* Amsterdam etc.: Elsevier on behalf of the International Cartographic Association (ICA).

Edsall, R. M. (2003). Design and usability of an enhanced geographic information system for exploration of multivariate health statistics. *Professional Geographer, 55*(2), 146-160.

Edsall, R. M. (2003). The Dynamic Parallel Coordinate Plot: Visualizing Multivariate Geographic Data Retrieved 2011, November 1, from http://www.geovista.psu.edu/publications/JSM99/paper.htm

Elwood, S. (2008). Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS. *GeoJournal, 72*(3), 173-183.

Ewing, R. M., & Cherry, J. M. (2001). Visualization of expression clusters using Sammon's non-linear mapping. *Bioinformatics (Oxford, England), 17*(7), 658-659.

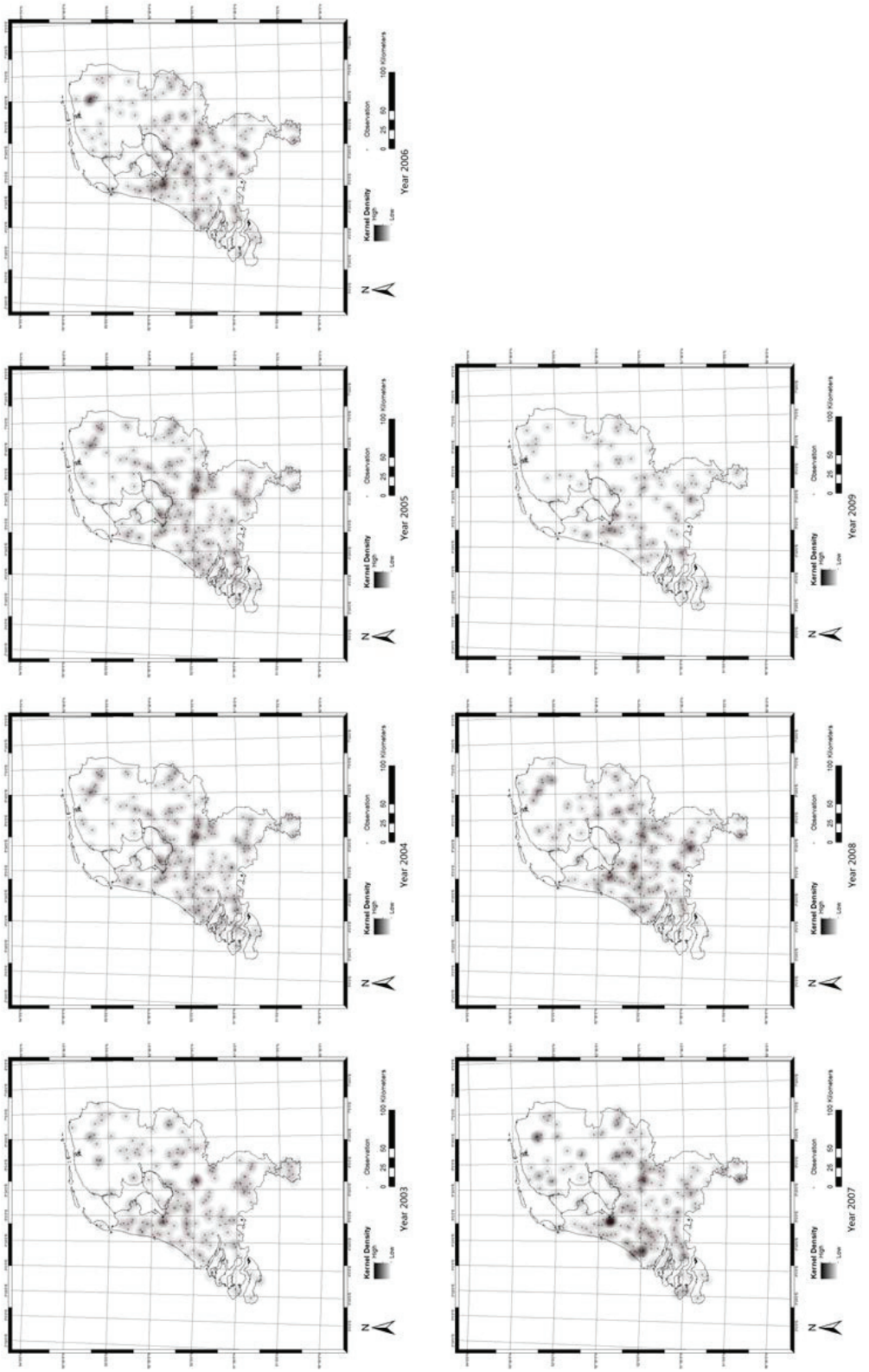
Fitter, A. H., & Fitter, R. S. R. (2002). Rapid changes in flowering time in British plants. *Science, 296*(GEOBASE), 1689-1691.
Fitzpatrick, M. C., Preisser, E. L., Ellison, A. M., & Elkinton, J. S. (2009). Observer bias and the detection of low-density populations. *Ecological Applications, 19*(GEOBASE), 1673-1679.
Freitas, L., & Bolmgren, K. (2008). Synchrony is more than overlap: measuring phenological synchronization considering time length and intensity. *Revista Brasileira de Botânica, 31*, 721-724.
Gardliner, L. (2009). What is Phenology? Retrieved 2011, November,11, from http://www.windows2universe.org/earth/climate/what_is_phenology.html
Goodchild, M. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal, 69*(4), 211-221.
Guo, D., Gahegan, M., MacEachren, A. M., & Zhou, B. (2005). Multivariate analysis and geovisualization with an integrated geographic knowledge discovery approach. *Cartography and Geographic Information Science, 32*(2), 113-132.
Ho Van, Q., Astrom, T., & Jern, M. (2009, 12-13 Oct. 2009). *Geovisual analytics for self-organizing network data.* Paper presented at the IEEE Symposium on Visual Analytics Science and Technology, 2009 (VAST 2009).
Hudson, I., Keatley, M., & Lee, S. (2011). Using Self-Organising Maps (SOMs) to assess synchronies: an application to historical eucalypt flowering records. *International Journal of Biometeorology*, 1-26.
Hudson, J. C., & Fowler, P. M. (1966). *The concept of pattern in geography* (First ed. Vol. 1): Department of Geography, University of Iowa.
Inselberg, A. (2009). Parallel Coordinates: Intelligent Multidimensional Visualization. In D. Plemenos & G. Miaoulis (Eds.), *Intelligent Computer Graphics 2009* (Vol. 240, pp. 123-141): Springer Berlin / Heidelberg.
Kaski, S. (1997). Data Exploration Using Self-Organizing Maps. *Acta Polytechnica Scandinavica, Mathematics, Computing and Management in Engineering Series No. ~82*.
Keatley, M. R., Hudson, I. L., & Fletcher, T. D. (2004). Long-term flowering synchrony of box-ironbark eucalypts. *Australian Journal of Botany, 52*(GEOBASE), 47-54.
Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., & Melançon, G. (2008). Visual Analytics: Definition, Process, and Challenges
Information Visualization. In A. Kerren, J. Stasko, J.-D. Fekete & C. North (Eds.), (Vol. 4950, pp. 154-175): Springer Berlin / Heidelberg.
Khalili, N., Wood, J., & Dykes, J. (2009). Mapping geography of social networks Retrieved 2011, Dec 29, from http://city.academia.edu/JasonDykes/Papers/616078/Mapping_the_Geography_of_Social_Networks
KNMI. (2004). PRISM standard name tables Retrieved 2011, December 30, from http://www.knmi.nl/~velthove/PRISM/CF/PRISM_standard_names_V1.3.html#recommendations
KNMI. (2011). The Royal Netherlands Meteorological Institute, from http://www.knmi.nl/
Koenig, W. D. (2006). Spatial synchrony of monarch butterflies. *American Midland Naturalist, 155*(1), 39-49.
Kohonen, T. (1984). *Self-organization and associative memory*. Berlin: Springe-Verlag.
Kohonen, T. (1995). *Self Organizing Maps*. Berlin: Springe-Verlag.
Kohonen, T. (2001). *Self-Organizing Maps* (Third extended ed.): Springe-Verlag.
Kraak, M. J., & Ormeling, F. J. (2010). *Cartography : visualization of geospatial data : also as e-book* (Third edition ed.). Harlow: Pearson Education.
Kwan, M.-P. (2004). GIS methods in time-geographic research: Geocomputation and geovisualization of human activity patterns. *Geografiska Annaler, Series B: Human Geography, 86*(GEOBASE), 267-280.
Liu, Y. G., Weisberg, R. H., & He, R. Y. (2006). Sea surface temperature patterns on the West Florida Shelf using growing hierarchical self-organizing maps. *Journal of Atmospheric and Oceanic Technology, 23*(2), 325-338.
Lundblad, P., Jern, M., & Forsell, C. (2008). *Voyage analysis applied to geovisual analytics*. Los Alamitos: Ieee Computer Soc.
MacEachren, A. M., & Kraak, M. J. (2001). Research challenges in geovisualization. *JournalCartography and geographic information science, 28*(1), 3-12.
Milla, R., Castro-Díez, P., & Montserrat-Martí, G. (2010). Phenology of Mediterranean woody plants from NE Spain: Synchrony, seasonality, and relationships among phenophases. *Flora - Morphology, Distribution, Functional Ecology of Plants, 205*(3), 190-199.

Miller-Rushing, A. J., Inouye, D. W., & Primack, R. B. (2008). How well do first flowering dates measure plant responses to climate change? The effects of population size and sampling frequency. *Journal of Ecology, 96*(GEOBASE), 1289-1296.
Nakaya, T., & Yano, K. (2010). Visualising Crime Clusters in a Space-time Cube: An Exploratory Data-analysis Approach Using Space-time Kernel Density Estimation and Scan Statistics. *Transactions in GIS, 14*(3), 223-239.
NASA , L. (2011a). MCD12Q1 Retrieved 2011, December 2, from https://lpdaac.usgs.gov/lpdaac/products/modis_products_table/land_cover/yearly_l3_global_500_m/mcd12q1
NASA , L. (2011b). MODIS Data Pool Retrieved 2011, December 2, from https://lpdaac.usgs.gov/lpdaac/get_data/data_pool
NPN. (2011a). About Phenology Retrieved 2011, August 17, from http://www.usanpn.org/about/phenology
NPN. (2011b). USA National Phenology Network Retrieved 2011, Octorber 20, from http://www.usanpn.org/
Peuquet, D. J. (2002). *Representations of space and time*. New York: The Guilford Press.
Post, E., Pedersen, C., Wilmers, C. C., & Forchhammer, M. C. (2008). Warming, plant phenology and the spatial dimension of trophic mismatch for large herbivores. *Proceedings of the Royal Society B: Biological Sciences, 275*(GEOBASE), 2005-2013.
RDocumentation. (2010). Sammon's Non-Linear Mapping Retrieved 2011, December 11, from http://127.0.0.1:18749/library/MASS/html/sammon.html
Rossi, M., da Silva Rodrigues, L., Ishino, M., & Kestring, D. (2011). Oviposition pattern and within-season spatial and temporal variation of pre-dispersal seed predation in a population of <i>Mimosa bimucronata</i> trees. *Arthropod-Plant Interactions, 5*(3), 209-217.
Seeger, C. J. (2008). The role of facilitated volunteered geographic information in the landscape planning and site design process. *GeoJournal, 72*(GEOBASE), 199-213.
Skupin, A., & Hagelman, R. (2005). Visualizing demographic trajectories with self-organizing maps. *Geoinformatica, 9*(2), 159-179.
Smith, M. J. d., Goodchild, M. F., & Longley, P. A. (2008). *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques, and Software Tools* (Second ed.): Blackwell Publishing Ltd.
Telmee. (2007). National databank Retrieved 2011,December,30, from http://telmee.nl/index.php?lan=112344
Tesquet, P. O. (2009). En Iran, "révolution Twitter" ou révolution tweetée? Retrieved 2011, October 25, from http://www.lexpress.fr/actualite/monde/proche-orient/en-iran-revolution-twitter-ou-revolution-tweetee_767535.html
Tulloch, D. (2008). Is VGI participation? From vernal pools to video games. *GeoJournal, 72*(3), 161-171.
van Vliet, A. J. H., de Groot, R. S., Bellens, Y., Braun, P., Bruegger, R., Bruns, E., . . . Sparks, T. (2003). The European Phenology Network. *International Journal of Biometeorology, 47*(4), 202-212.
Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *Neural Networks, IEEE Transactions on, 11*(3), 586-600.
Vries, H. d., & Verheul, D. (2006). Waarneming, from http://waarneming.nl/index.php?lang=en&local=nl
Wageningen UR, & VARA. (2001). De Natuurkalender Retrieved 2011, May 29, from http://www.natuurkalender.nl/
Wanstreeta, C. E., & Steina, D. S. (2011). Presence Over Time in Synchronous Communities of Inquiry. *American Journal of Distance Education Volume 25*(Issue 3).
Wehrens, R., & Buydens, L. M. C. (2007). Self- and Super-organizing Maps in R: The kohonen Package. *Journal of Statistical Software, 21*(5), 1--19.

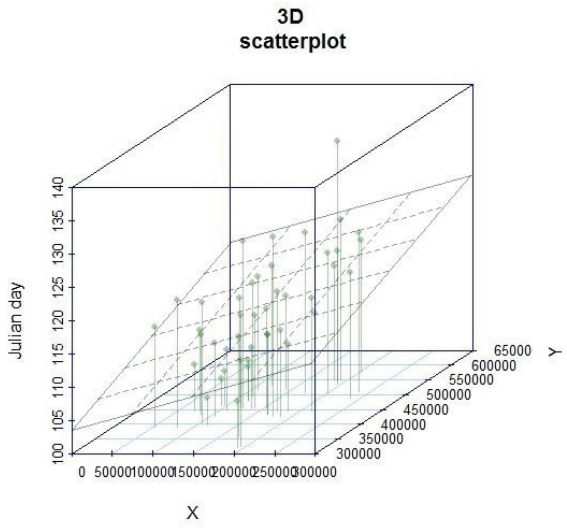
APPENDICES

Appendix A

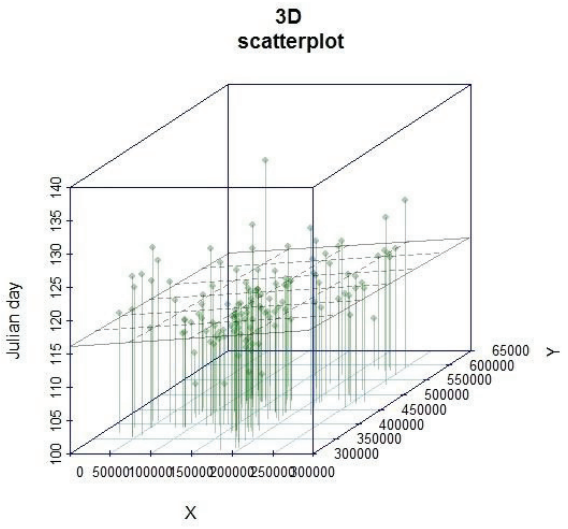
(1) Geographical density maps for Gierzwaluw (swift) through 7 years (2003 to 2009)



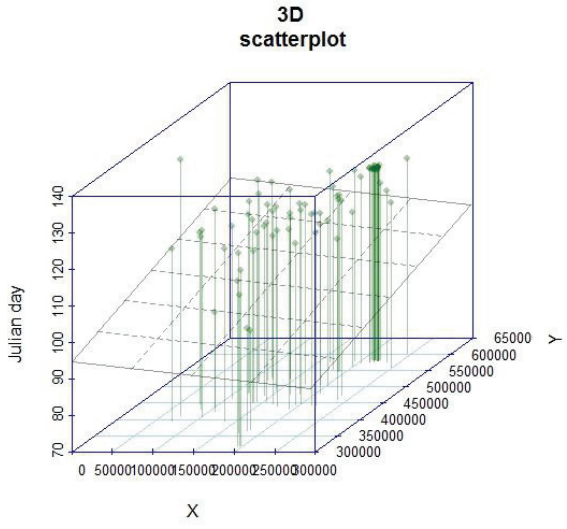
(2) 3D scatterplot for species (tree paardekastanje & swift bird gierzwaluw) from 2003 to 2009 and a linear regression (Julian_day over x + y) plane for study overall spatiotemporal pattern



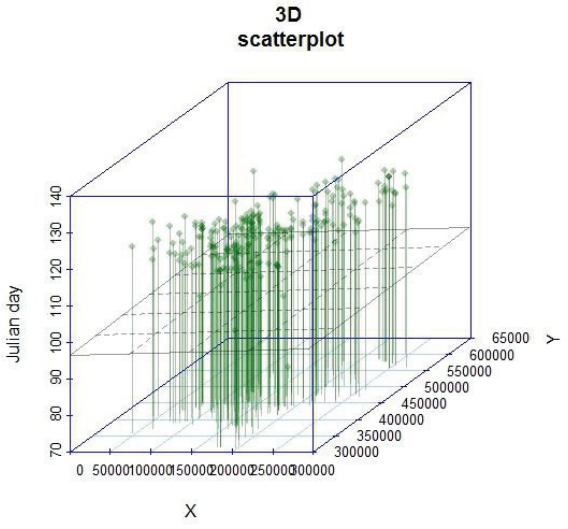
Paardekastanje (witte), Year 2003, multiple R-squared: 0.3043



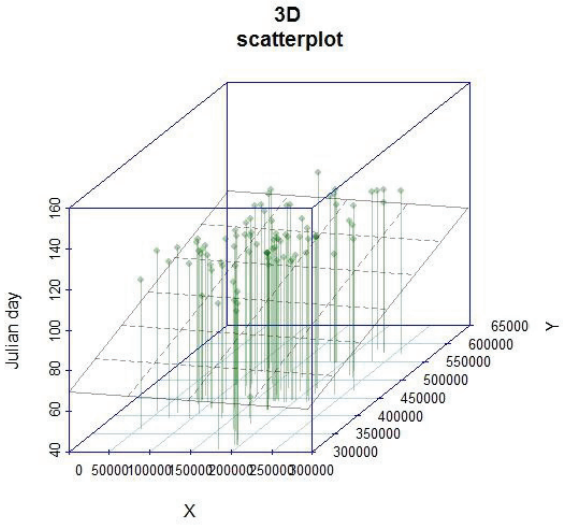
Gierzwaluw, Year 2003,multiple R-squared: 0.0189



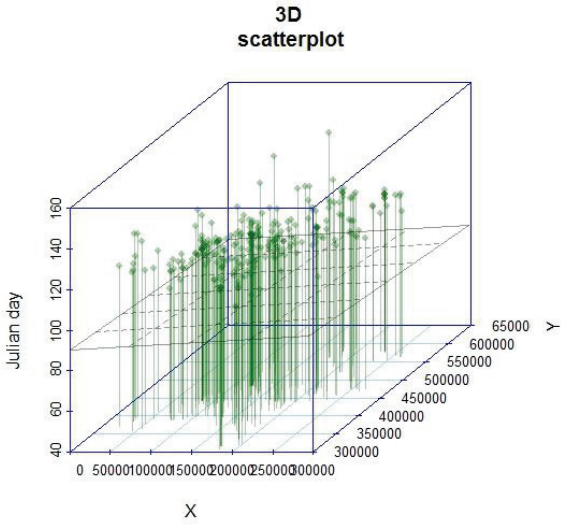
Paardekastanje (witte), Year 2004, multiple R-squared: 0.1945



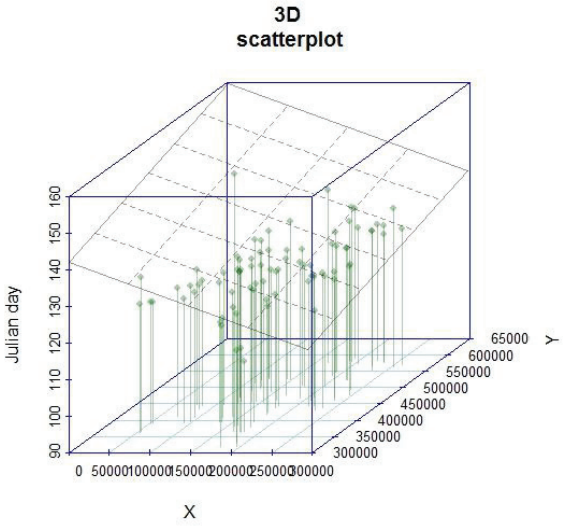
Gierzwaluw, Year 2004,multiple R-squared: 0.01312



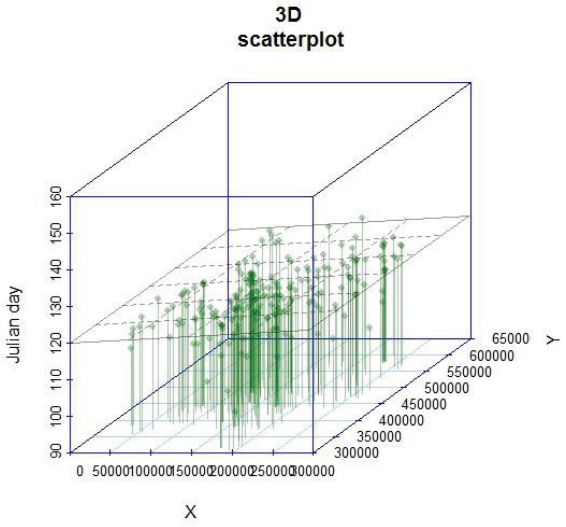
Paardekastanje (witte), Year 2005, multiple R-squared: 0.1066



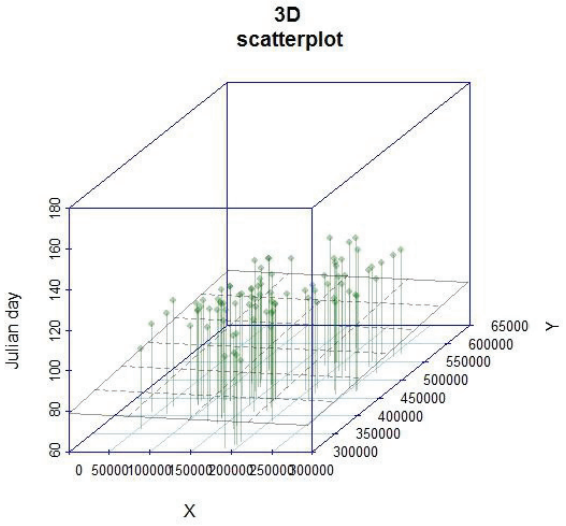
Gierzwaluw, Year 2005,multiple R-squared: 0.007801



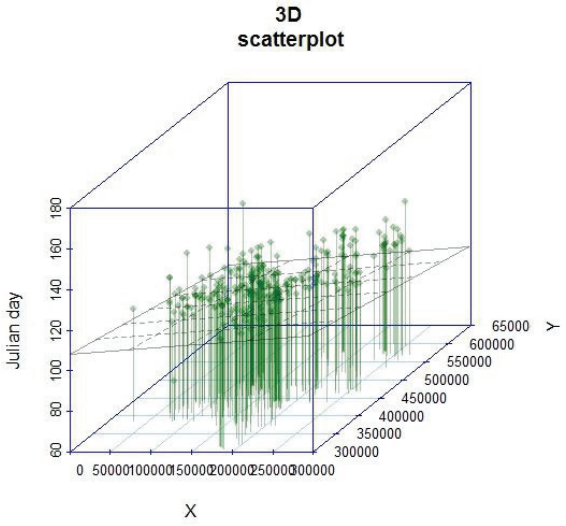
Paardekastanje (witte), Year 2006, multiple R-squared: 0.03942



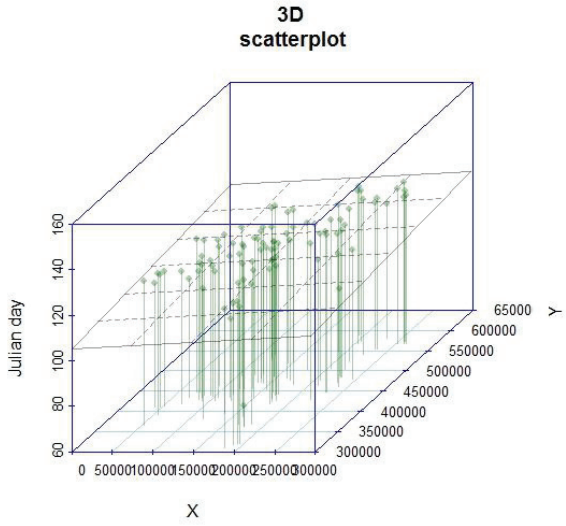
Gierzwaluw, Year 2006,multiple R-squared: 0.01117



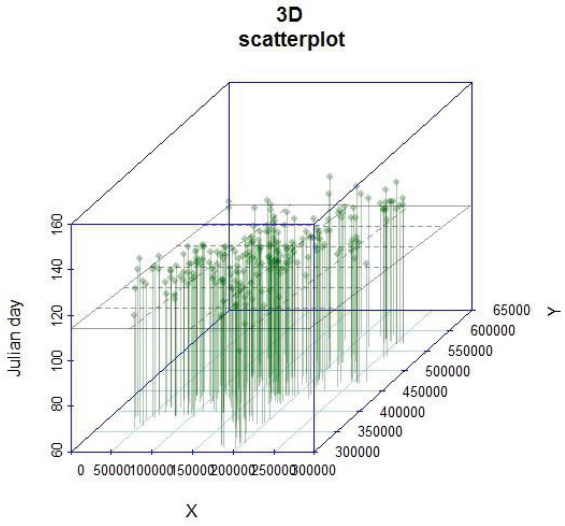
Paardekastanje (witte), Year 2007, multiple R-squared: 0.05734



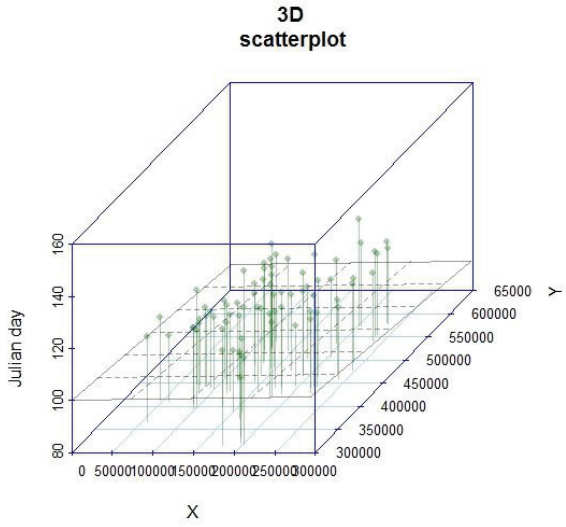
Gierzwaluw, Year 2007,multiple R-squared: 0.01391



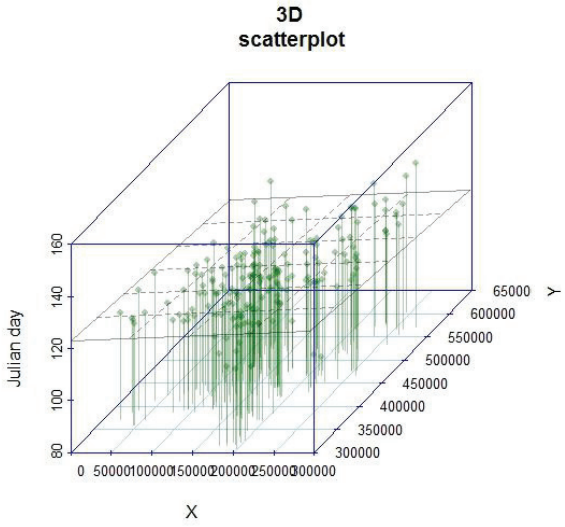
Paardekastanje (witte), Year 2008, multiple R-squared: 0.06878



Gierzwaluw, Year 2008,multiple R-squared: 0.00549



Paardekastanje (witte), Year 2009, multiple R-squared: 0.05398



Gierzwaluw, Year 2009,multiple R-squared: 0.01488

Appendix B

R codes:

(1) Subset data and create new dataset

```
## input dataset
> mydata<-read.csv(file.choose())
##creat new data frame
> df.s<-data.frame(x=mydata$x, y=mydata$y, z=mydata$Julian_day,
a=mydata$species,b=mydata$Year, c=mydata$T5_sum, d=mydata$precip_sum,e=mydata$EV_sum)
#subset by species and year
> df.p<- subset (df.s, a=="Paardekastanje (witte)")
> df.p05<- subset (df.p, b=="2005")
```

(2) Establish space-time-cube and three dimensional presentation

```
## load scatterplot3d package
> library(scatterplot3d)
## input dataset
> mydata<-read.csv(file.choose())
## define axes
> y<- df.p05$y
> x<- df.p05$x
> t<- df.p05$z
> c<- df.p05$c
> d<- df.p05$d
> e<- df.p05$e
## space-time-cube
> s3d <- scatterplot3d
(x,y,c,type="h",xlim=c(15000,276000),ylim=c(304000,612000),zlim=c(150,1100),angle=30,pch=16,ce
x.symbols=0.8,color= rgb (0,100,0,90,maxColorValue=255), xlab="X",ylab="Y",zlab="above 5
degree Temperature sum",col.axis="blue", col.grid="lightblue",cex.axis=0.7)
# Now adding a regression plane to the "scatterplot3d"
> fit <- lm(t ~ x+y)
> s3d$plane3d(fit,lty.box = "solid",col= rgb (0,0,0,100,maxColorValue=255))
> title("3D scatterplot year 2005")
```

(3) Establish parallel coordinates plots

```
## load rggobi package
```

```
> library(rggobi)
> mydata<-read.csv(file.choose())
## input data in ggobi interface
> h <- ggobi (mydata)
```

(4) Establish SOM

```
> library(kohonen)
> mydata<- read.csv(file.choose())
##set random sampling seed
> set.seed(12)    ## create SOM
> mydata.som <-som(scale(mydata[,2:17]),grid=somgrid(3,4,"hexagonal"),rlen=1000000)
##decied the cluster number with Sammon function
> mydata.sam<- sammon(dist(mydata.som$codes))
> plot (mydata.sam$points,xlab="x", ylab="y",main="Sammon's Non-Linear Configuration
Mapping",cex=2)
> text (mydata.sam$points,as.character(1:nrow(mydata.som$codes)),cex=0.8,col="darkgreen")
>  legend ( "topleft","neuron", pt.cex=2, bty="n",pch=1 )
## plot SOM
> plot (mydata.som, type="mapping", labels= mydata$species ,main="mapping plot")
## use hierarchical clustering to cluster the codebook vectors
## cut the tree into 5 clusters and reconstruct the upper part of the tree from the cluster centers.
> som.hc <- cutree (hclust(dist(mydata.som$codes)),5)
> add.cluster.boundaries(mydata.som,som.hc,col="red")
```

(5) Establish boxplot and bar plot

```
> mydata<- read.csv(file.choose())
## Barplot
> barplot ( mydata ,main="Total Observations by Landcovers and Years", xlab="Year",
ylab="Total",col=c("darkgreen","red","lightgreen","grey"),space=0.7)
> legend ( "topleft",rownames(mydata),fill=c("darkgreen","red","lightgreen","grey"), cex=0.8,
bty="n" )
##Box plot
> mydata2<- read.csv(file.choose())
> boxplot (Julian.day~Year,data=mydata2,main="boxplot of
swift(Gierzwaluw)",xlab="Year",ylab="Julian_day",space=0.7,col=rainbow(8) )
```